University of Zurich[UZH]

Department of Informatics

Understanding the Impacts of Offline and Online Social Influence on Open Source Software Project Success

Dissertation submitted to the
Faculty of Business, Economics and Informatics
of the University of Zurich

to obtain the degree of
Doktor der Wissenschaften, Dr. sc.
(corresponds to Doctor of Science, PhD)

presented by
Xiao Li
from China

approved in April 2018

at the request of
Prof. Dr. Daning Hu
Prof. Dr. Jiuyong Li

The Faculty of Business, Economics and Informatics of the University of Zurich hereby authorizes the printing of this dissertation, without indicating an opinion of the views expressed in the work.

Zurich, April 4, 2018

The Chairman of the Doctoral Board: Prof. Dr. Sven Seuken

Abstract

Open source software development (OSS) has become a great success which attracted attentions from both the academia and industry. With the prevalence of social networking-like websites such as, Open Hub, GitHub and Source Forge, etc. OSS has provided new ways for software firms and developers to interact and collaborate with each other to create innovative software products. The success of an OSS project often largely relies on the voluntary participations, contributions, and efficient collaborations among its team members. Therefore, it is important for both researchers and OSS stakeholders to systemically understand the mechanisms which affect OSS project success thus develop strategies or tools for OSS development management.

Previous OSS project success research mainly focused on OSS developers' motivations of project participations and contributions. Since the successful development of an open source software (OSS) project requires a steady supply of motivated software developers. However, enabled by successful social coding websites, OSS developers may engage in both offline (e.g., face-to-face) and online interaction ways (e.g., social networking services such as microblogging) to collaborate with each other. The effective collaboration and interaction among OSS developers are important for creating successful software products. Consequently, investigating the OSS project success with a social influence perspective can provide us insights on how developers are influenced by each other's voluntary behaviors rather than incentive-based behaviors which is common in traditional software development model in firms. Based on such insights, the thesis contributes a rich understanding of the impacts of both offline (e.g., face-to-face) and online social interactions enabled by the social networking services on the OSS project success. More specifically, this thesis includes two main studies as contributions.

The study I investigates the impacts of OSS project members' geographic dispersions on their team performance (project success). Existing empirical studies focus on analyzing the impacts of either virtual or face-to-face collaborations alone, but rarely studied the situation when both settings coexist in a heavily technology-dependent context like OSS development. Moreover, the impacts of spatial and temporal geographic dispersions were not clearly distinguished before. To address these, we use the instrument variable estimation to analyze data from a real-world online OSS community. The results show that the geographic distances among team members negatively affect project success, even after controlling temporal dispersion and other related factors. The findings provide insights for OSS project managers to help devise strategies and policies to improve team performance and thus project success.

The study II empirically examines the impacts of word of mouth (WOM) and observational learning (OL) on OSS developers' initial and sustained participation behaviors, using data from a large OSS platform-Open Hub. Since, open source software (OSS) development platforms are increasingly using social networking-like functions such as microblogging, aiming to use developers' social influence to attract more high-quality project participation. However, social influence is largely overlooked in OSS participation research and has often been studied from an economic utility framework in existing literature. Such a framework may be not suitable for analyzing the non-monetary motivations behind OSS developer participations. The preliminary results show that the online social influence has significant but rather different impacts on initial and sustained OSS project participation. Specifically, the impacts of WOM on developers' sustained participation faded away after initial participation as they can better evaluate the underlying project and its members' opinion.

To my best of knowledge, this thesis is the first study that investigates the mechanisms which affect the success of OSS projects with both offline and online social influence perspectives.

Moreover, the thesis provides causal interpretations for the findings with rigorous econometric methods. The findings of the thesis not only empirically validate the theories for OSS research but also provide practical insights for OSS stakeholders.

Zusammenfassung

Open-Source-Software-Entwicklung (OSS) ist zu einem großen Erfolg geworden, der sowohl von der Wissenschaft als auch von der Industrie Beachtung fand. Mit der Verbreitung von Social-Coding-Websites wie Open Hub, GitHub und Sourceforge.net, etc., hat OSS neue Möglichkeiten für Softwarefirmen und -entwickler geschafft, mit denen sie interagieren und zusammenarbeiten können, um innovative Softwareprodukte zu entwickeln. Der Erfolg eines OSS-Projekts hängt oft weitgehend von freiwilligen Beteiligungen, Beiträgen und effizienten Zusammenarbeiten zwischen seinen Teammitgliedern ab. Daher ist es wichtig, dass sowohl die Forscher als auch die OSS-Stakeholder die Mechanismen verstehen, die die Leistung des OSS-Teams beeinflussen, und Strategien oder Werkzeuge zur Verbesserung des Projekterfolgs entwickeln.

Frühere Erfolgreiche studien von früher zum OSS-Projekt konzentrierten sich hauptsächlich auf die Motivation von OSS-Entwicklern von Projektbeteiligungen und -beiträgen. Seit der erfolgreichen Entwicklung eines Open-Source-Software (OSS) -Projekts bedarf es eines stetigen Angebots an motivierten Softwareentwicklern. Jedoch können die OSS-Entwickler, vermittelt durch die erfolgreichen Social-Coding-Websites, sowohl offline (z.B. Face to Face) als auch Online-Interaktionsweisen (z.B. soziale Netzwerkdienste wie Microblogging) kooperieren, um miteinander zusammenzuarbeiten. Folglich ist die effektive Zusammenarbeit und Interaktion zwischen den OSS-Entwicklern ausschlaggebend für die Erstellung eines erfolgreichen Softwareprodukts. Daher kann die Studien des OSS-Projekterfolgs mit einer Perspektive des sozialen Einflusses uns Einblicke darüber geben, wie Entwickler durch das freiwillige Verhalten der anderen beeinflusst werden, anstatt durch anreizbasierte Verhaltensweisen, wie es im traditionellen Softwareentwicklungsmodell in Unternehmen üblich ist. Basierend auf solchen Einsichten trägt die Doktorarbeit zu einem umfassenden

Verständnis der Auswirkungen von sowohl Offline- (z.B. Face-to-face) als auch Online-sozialen Interaktionen bei, die durch die Social-Networking-Dienste zum Erfolg des OSS-Projekts ermöglicht werden. Im Einzelnen umfasst die Arbeit hauptsächlich zwei Studien, die zur Erfolgsliteratur des OSS-Projekts beitragen.

Die erste Studie untersuchte die Auswirkungen der geografischen Aufteilung von OSS-Projektmitgliedern auf ihre Teamleistung (OSS-Projekterfolg). Bestehende empirische Studien konzentrierten sich auf die Analyse der Auswirkungen von virtuellen oder direkten Kollaborationen allein, untersuchten jedoch selten die Situation, in der beide Einstellungen in einem stark technologieabhängigen Kontext wie der OSS-Entwicklung koexistieren. Darüber hinaus waren die Auswirkungen räumlicher und zeitlicher geographischer Unterschiede bisher nicht klar voneinander abgegrenzt. Um diese zu adressieren, wurde die Instrumentenvariablenschätzung verwendet, um Daten von einer realen Online-OSS-Gemeinschaft zu analysieren. Die Ergebnisse zeigten, dass die geografische Entfernungen zwischen den Teammitgliedern den Projekterfolg negativ beeinflussen, selbst nach Kontrolle der zeitlichen Streuung und anderer damit verbundener Faktoren. Die Ergebnisse liefern Einblicke für OSS-Projektmanager, um Strategien und Richtlinien zur Verbesserung der Teamleistung und damit des Projekterfolgs zu entwickeln.

Die zweite Studie untersuchte empirisch die Auswirkungen von word of mouth (WOM) und observational learning (OL) auf das initiale und anhaltende Partizipationsverhalten von OSS-Entwicklern unter Verwendung von Daten einer großen OSS-Plattform - Open Hub. Seither verwenden Open-Source-Software (OSS) -Entwicklungsplattformen zunehmend soziale Netzwerk-ähnliche Funktionen wie Microblogging, mit dem Ziel, den sozialen Einfluss von Entwicklern zu nutzen, um mehr Partizipation von qualitativ hochwertigen Projekten zu gewinnen. Der soziale Einfluss wird jedoch in der OSS-Partizipationsforschung weitgehend übersehen und wurde oft in der bestehenden Literatur aus der Perspektive eines ökonomischen Nutzenrahmens untersucht. Ein solcher Rahmen ist möglicherweise nicht geeignet, um die oft nicht-monetären Beweggründe für OSS-Entwicklerbeteiligungen zu analysieren. Die vorläufigen Ergebnisse zeigten, dass der soziale Einfluss im Internet signifikante, aber unterschiedliche Auswirkungen auf die initiale und anhaltende OSS-Teilnahme hat. Konkret gingen die Auswirkungen von WOM auf die nachhaltige Teilnahme von Entwicklern nach der ersten Teilnahme zurück, da sie das zugrunde liegende Projekt und die Meinung seiner Mitglieder besser bewerten können.

Nach meines Wissens ist diese Doktorarbeit die erste Studie, die die Mechanismen untersucht, die den Erfolg von OSS-Projekten sowohl aus der Offline- als auch aus der Online-Sicht beeinflussen könnten. Die Arbeit liefert auch kausale Interpretationen für die Ergebnisse mit strengen ökonomischen Methoden. Die Ergebnisse der Arbeit validieren nicht nur empirisch die Theorien in OSS-Literaturen, sondern bieten auch praktische Einblicke für OSS-Stakeholder.

Acknowledgements

This thesis would not be possible without the support and help of my supervisors, PhD committee, colleges, friends and my parents.

First of all, I would like to thank my supervisor Prof. Dr. Daning Hu. He has always been a very professional researcher and supervisor. It is him who help me to conduct the research. Moreover, he gave me a lot of support and confidence which made the journey of my PhD study not only substantive but also promising. I am very grateful for his patience and generosity to share rich experiences concerning both research and the way of life.

My appreciation also goes to Prof. Dr. Jiuyong Li who served as the co-advisor of my PhD thesis. The constructive comments and valuable suggestions for my PhD thesis offered by him considerably helped me in terms of improving the quality of this thesis.

I want to thank Prof. Dr. Michael Böhlen not only for serving as the chairman of the doctoral board for this dissertation but also for providing constructive comments and valuable suggestions for my PhD thesis proposal.

My gratitude has to go to Prof. Dr. Gerhard Schwabe as well. Thanks for your constructive comments and valuable suggestions of my PhD thesis proposal. The constructive comments about the PhD thesis proposal provided by him make my PhD thesis not only academically promising but also practically interesting.

I also extend my gratitude to my research collaborators: Professor Xiaoquan (Michael) Zhang, Professor Harry Jiannan Wang, Dr. Xuan Yang, Dr. Jiaqi Yan, Dr. Ji Wu, and Dr. Jiaoyan Chen. It is difficult to put into words how deeply grateful I am for our collaboration and friendships. I have learned so much from you about research and work. Without the creative ideas and experiences of you, the research in this thesis would not be possible.

My thanks also go to the staff at the Department of Informatics, University of Zurich: Monika Lauber, Barbara Jost, Claudia Leibundgut, Beat Rageth, Nathan Labhart, Daniela Bärtschi, Enrico Solca. Thank you for your help in the last four years and making my life and work so much easier and enjoyable.

Last but not least, I thank for my parents who gave me invaluable encouragements and supports. Without the love from my parents, none of the possibilities mentioned above can come true.

This thesis is dedicated to my parents

Table of Contents

Part 1: Synopsis

Understanding the Impacts of Offline and Online Social Influence On Open Source Software Project Success

Xiao Li
Department of Informatics
University of Zurich

List of Figures

List of Tables

Chapter 1: Introduction

Nowadays, open source software (OSS) development has become a great success which attracted millions of voluntary developers who collaborated with one another in various project teams to produce software products like Linux, Firefox, and MySQL. Krogh, Haefliger, et al. (2012) defined OSS as “software where users inspect the source code, modify it, and redistribute modified or unmodified versions for others to use”. With the fast development of online social coding websites, such as Open Hub, GitHub and Source Forge, etc., such an approach has provided new ways for software firms and individuals to interact and collaborate with their customers in software development and maintenance (Di Gangi and Wasko 2009) and has been supported by more and more major software vendors such as IBM, Microsoft, Google, etc.

Consequently, information system (IS) managers have become increasingly relied on voluntary developers outside firms who are not able to influence through traditional management practices within firms such as pay incentives and output-based control. The success of an OSS project often largely relies on the voluntary participations, contributions, and efficient collaborations among its team members. Therefore, it is crucial for IS researchers and practitioners to understand the factors that affect OSS team performance and thus develop strategies or tools for improving project success.

Previous OSS studies focused on developers’ motivations of project participations and contributions. This is understandable since the successful development of an OSS project requires a steady supply of motivated software developers. But such studies mainly adopted the intrinsic/extrinsic motivation framework in the self-determination theory (SDT) (Deci and Ryan 1985). The intrinsic motivations for participations are the factors related to OSS developers' needs for satisfying themselves such as altruism and enjoyment, while extrinsic motivations are usually derived from external rewards such as desire for good reputation and career advancement opportunities (Hertel et al. 2003; Krishnamurthy 2006).

However, while these studies inform us well on how OSS developers' intrinsic and extrinsic motivations, little is known about why they achieve project success (i.e., produce high-quality software) when they do. Von Krogh et al. (2012) suggested that new perspectives other than the dominating SDT are needed to study the success of OSS projects.

Recent years scholars have increasingly shown interests in the impacts of social influence, defined by Sinan Aral (2011) as "how the behaviors of one's peers change the utility one expects to receive from engaging in a certain behavior and thus the likelihood that (or extent to which) one will engage in that behavior", especially in marketing and social network research (Aral et al. 2009; Aral and Walker 2011; Aral and Walker 2012; Bandura 1971; Dodds and Watts 2004; Katz and Lazarsfeld 1966). OSS development largely relies on online and offline communities where developers (Hippel and Krogh 2003) collaborate, exchange ideas and resources, and thereby form various social networks. Currently, more and more developers use various social networking services such as microblogging and tagging to collaborate online. For instance, an OSS project leader who is a well-known developer may promote her project in her microblog and consequently influence her followers' decisions in participating or contributing to that project, thereby achieving better project success. Therefore, it is critical to systematically study how social influence among OSS developers affect their team performance (project success). This thesis proposes a relational (social influence) perspective that focuses on studying the impacts of social influence among OSS developers on their development behaviors and subsequent project success (team performance).

The structure of the cumulative thesis is illustrated in the Figure 1. The thesis consists of three parts. Part 1 is the synopsis of this thesis. The synopsis summarizes the two studies which are contributions of the thesis by providing introduction, literature review, problem statement, research design, overall results, limitations and future research, and conclusion. Part 2 is the study I which contributes to the thesis. The study I investigates the impacts of the offline social influence on open source software (OSS) project success. Specifically, it investigates the impacts of geographic dispersion on OSS team performance which is indicated by the average rating score of the project. Part 3 is the study II which contributes to the thesis. As OSS development platforms are increasingly involving social networking-like functions such as microblogging, aiming to use developers' online social influence to attract more high-quality project participation. The online social influence among OSS developers enabled by the social networking technologies will influence their participation behaviors which are important to OSS project success. Thus, the study II examines the impacts of two types of online social influence: word of mouth (WOM) and observational learning (OL) on OSS developers' initial and sustained participation. As often with the cumulative thesis, this synopsis is partly redundant with the two papers presented as main contributions to the thesis.

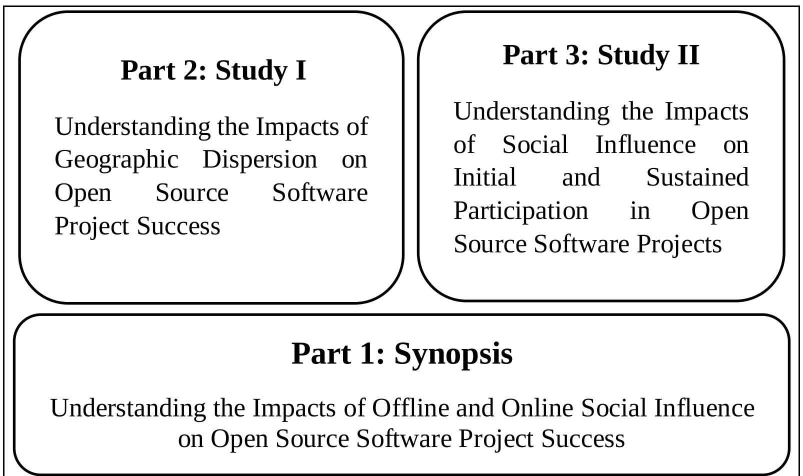

Figure 1: The Structure of the Thesis

Chapter 2: Research Background

2.1 OSS Project Communities

Open source software (OSS) development approach has been extensively studied in recent years (Raymond 2001; Roberts et al. 2006; Scacchi et al. 2006). Most previous studies aim to identify the factors influencing developers' motivation to contribution and its relations with project success (Hertel et al. 2003; Krishnamurthy 2006; Roberts et al. 2006; Subramanyam and Xia 2008). In general, the motivations can be categorized into two types – intrinsic and extrinsic motivations (Krishnamurthy 2006). The intrinsic motivations are the factors related to OSS developers' needs for satisfying themselves such as altruism and enjoyment, while extrinsic motivations are usually derived from external rewards such as desire for good reputation and career advancement opportunities.

Another main motivations discovered in previous OSS studies and central for our research is the developers' desire to gain reputation or good impressions from their peers. The lack of traditional monetary reward in an OSS development environment has made the reputation-based motivation as a major drive for developers' participations and contributions (Subramaniam et al. 2009). One of the most cited studies on OSS participation motivations done by Roberts et al. (2006), has examined the relationship among developers' participation, performance, and motivation using empirical data from the Apache project. They found that Apache developers' desire to gain high community reputation can lead to above average participation levels. Moreover, developers with higher status (i.e. better reputation) within the Apache project community are found to have significantly higher wages. Singh and Phelps (2012) adopted a social influence perspective to study the license choice of a new OSS project. Their findings suggested that the most important factor determining the license choice of a new project is the type of license chosen by existing projects that are socially closer to it in its inter-project network.

To summarize, these empirical findings suggested that social influence among OSS developers within their social network communities can be a potential driver of their development (participation and contribution) behaviors. In other words, OSS developers may be influenced by their friends' development behaviors since they cared about how they are perceived by the friends in their social network communities, in order to gain reputations or other benefits. However, previous OSS research largely overlooked the social influence perspective and mainly focused on intrinsic and extrinsic perspective.

2.2 Social Influence and OSS Project Success

Recent years there is an emerging trend of research which focuses on studying social influence, particular in marketing and social network analysis domains. Sinan Aral (2011) formally defined peer influence as "how the behaviors of one's peers change the utility one expects to receive from engaging in a certain behavior and thus the likelihood that (or extent to which) one will engage in that behavior". Social influence is not part of the self-determination. According to previous literature (Chatzisarantis et al. 2007), social influence refers to the phenomenon that other people's behaviors can influence the subject's behaviors, not through the internalization process. The main driver of the subject's behavior change is from the outside force such as persuasion, instead of the internalized motivations like the self-determination theory (SDT) (Deci and Ryan 1985) suggested.

There are mainly two types of studies for modeling and analyzing social influence among individuals in their social networks. The first type of studies investigates the social influence by flexibly modeling the correlations that exist in the observed choices or behaviors of individuals within a reference group (often within same geographical locations). For instance, the offline social interactions can influence neighborhoods peer's behaviors and outcomes effects (Choi et al. 2010; Dietz 2002; Sessions 2010; Yang and Allenby 2003). However, the online social interactions mediated by the internet have drawn great attention from researchers (Chen et al. 2011). There are two main types of online social influence studies examined in previous studies: word of mouth (WOM) and observational learning (OL) (Arndt 1967; Bandura 1977; Chen et al. 2011).

OSS developers often have a common non-profit goal – creating better open source software product rather than consumer communities which primarily focus on economic utility. OSS developers may engage in both offline (e.g., face-to-face) and online interaction ways (e.g., social networking services such as microblogging) to collaborate with each other. Therefore, studying social influence in OSS communities can provide us theoretical insights on how developers are influenced by each other's voluntary behaviors rather than incentive-based behaviors which is common in traditional software development model in firms.

2.2.1 Offline Social Influence

The peer and neighborhoods effects among individuals have been extensively explored in both marketing and organization science literatures (Choi et al. 2010; Dietz 2002; Sessions 2010). Studies of social influence in this type specify correlation structures such that behaviors or choices by individuals near one another in terms of geographic dispersions will generate similar outcomes. For instance, Yang and Allenby (2003) study the decisions of consumers to buy foreign or domestic automobiles. The idea is that consumers in certain regions may influence one another in terms buying cars, resulting in correlations of purchase choices. Such correlations may also be the result of individuals' desires to buy similar autos, or the result of common unobserved beliefs, such as patriotism. Dietz (2002) conducted the interdisciplinary research of neighborhood effects and argued that neighborhood effects are community influences on individual social or economic outcomes.

Currently, research in virtual teams has mostly explored online virtual communities coupled with neighborhoods or other spatially co-located communities (Hampton and Wellman 2003; Hampton 2007). Their findings suggested that the Internet, combined with a local online discussion group, transforms and enhances neighboring among individuals located in a suburb. Bell and Song (2007) examined customer trials at an online shopping website Netgrocer.com. Their research has drawn on studies in marketing and economics conjecture exposure spatially to proximate others (through direct social interaction or observation), can influence decisions of those who have yet to try. Their findings suggested that the neighborhood effect is significantly positive and economically meaningful. Scholars in marketing and social network studies have provided an indirect link between the geographic dispersion and social influence- the proximity in terms of geographic location among individuals will increase the social proximity thus facilitate the likelihood of social influence among them (Hagerstrand 1967; Jannik Meyners 2017; Levy and Goldenberg 2014). OSS project communities are regarded as typical virtual teams where the offline social relationships among developers can be formed and influence their interactions and collaborations, hence affect the OSS project success. It is important for us to study the impacts of offline social influence in terms of geographic dispersion on the OSS project success.

2.2.2 Online Social Influence

The other type of approaches aims to model social influence by using various explicit social relationships. Through our daily experiences, individuals are influenced by their family, colleagues, or friends in their social networks. Moreover, the increasing recognition of the role of social influence in online communities (e.g. social networking sites such as Facebook) has spurred renewed interests in modeling and understanding social influence through explicit social relationships among agents (Hartmann et al. 2008).

Social influence has been defined as “how the behaviors of one’s peers change the utility one expects to receive from engaging in a certain behavior and thus the likelihood that one will engage in that behavior” by (Aral 2011). It has been extensively studied in the economics and marketing literature (Arndt 1967; Bandura 1971; Katz and Lazarsfeld 1966). But these studies often employ an economic utility framework with cost-benefit analysis to examine the impacts of social influence on one’s purchase decisions.

However, OSS developers’ participations are usually motivated beyond pure economic considerations, such as altruism, learning, and desire to gain community reputation (Hertel et al. 2003; Raymond 2001; Stewart 2005). Therefore, social influence may have quite different impacts on OSS project participations than the previously well-studied consumer adoption behaviors in the marketing and economic literature. However, to the best of my knowledge, little research has explored the impacts of social influence on OSS project participations. Prior OSS studies also suggested that project success largely relies on two types of developers’ participation behaviors - initial participation and sustained participation (i.e., continuous contribution) (Fang and Neufeld 2009; Roberts et al. 2006). Thus, this thesis aims to explore the impacts of online social influence on the initial and sustained participation in OSS projects.

In addition, this thesis also relates to the virtual team research from two perspectives. First, from the theoretical perspective, the previous studies of the media richness theory mainly focus on the media choices. The main gap those online media choices studies is surveying the media choice of online message senders not by examining the actual performance effects of media use (Dennis and Kinney 1998). Second, from the practical perspective, OSS project leaders largely rely on the online social-networking communities (e.g., social networking services such as microblogging) to efficiently manage their teams for achieving project success (Yang et al. 2013). OSS developers acquire and allocate diverse problem-solving skills from fellow developers through their participation in software development via Internet-based social network (Fong Boh et al. 2007). Allocating knowledge resources through the virtual team help developers to design better software, anticipate potential problems, solve thorny problems, and better identify user needs during development, which are essentially crucial for project OSS success.

2.3 The Impacts of Geographic Dispersion on OSS Team Performance

2.3.1 Spatial vs. Temporal Geographic Dispersion

Based on previous studies on OSS and virtual teams (Fang and Neufeld 2009; O'Leary and Cummings 2007; O'Leary and Mortensen 2010), the thesis mainly examines two dimensions of geographic dispersion among OSS team members: spatial and temporal. Previous studies mainly focused on the spatial geographic dispersion measured by the average geographic distance among team members. They have found a positive association between physical proximity and interpersonal communication efficiency (Festinger et al. 1950; Kiesler and Cummings 2002)

On the other hand, it was also found that temporal dispersion is negatively correlated with team performance (Griffith et al. 2003; Martins et al. 2004). A team's temporal geographic dispersion is often measured by the time zone differences among team members (O'Leary and Cummings 2007). Team members who are distributed in distant time zones, usually share fewer work hours and need to put more efforts in synchronize their working time with others considerably. Because of the lack of synchronous collaboration or the high costs for doing so, greater temporal dispersion often hinders real-time problem solving and spontaneous communications (Burke et al. 1999; Dennis et al. 1988). The difficulties of synchronize communications grow as the temporal dispersion increases among team members (Davison et al. 2006).

Then another major research challenge is to distinguish the impacts between spatial geographic dispersion and temporal geographic dispersion on team performance (OSS project success). This is mainly because in real world many team members are dispersed both in the small spatial and temporal distances (O'Leary and Cummings 2007). In such situations, it is difficult to find out the good performance of a team is due to face-to-face collaboration enabled by close physical proximity or enabled by being in the same time zone.

2.3.2 Geographic Dispersion and OSS Teams

This research categorizes the geographic dispersions of OSS team members into two settings: geographically co-located and distributed ones. Geographically co-located team members often reside or work in the same geographical location. The level of location often depends on the lens of the observer and the application context. It could be a building, a street or a city. In the context of OSS development, we generally consider OSS team members who are within the same city as geographically co-located. By this definition, such co-located team members are in closer spatial proximity with each other than geographically distributed OSS team members (i.e., developers who are not in the same city). Thus, geographically co-located OSS team members in general more likely to have face-to-face interactions than distributed members, therefore are more likely to influence each other's development behaviors.

In most OSS developer teams, both settings often co-exist, effectively creating three situations. When the majority team members are in the same location (city), this team can be in general considered as a co-located team (Sharp et al. 2012). An extreme example can be a group people work in the same level of a building. On the other hand, if the majority are distributed in multiple locations (cites), it can be considered as a distributed team. An extreme case can be a small group of developers without any two of them are in the same country. However, previous research on OSS (Crowston et al. 2005; Crowston and Scozzi 2002) and global software development (GSD) teams (Jalali and Wohlin 2012; Šmite et al. 2010) found that most such teams have a mixture of both settings of members and can be called hybrid teams (Staples and Webster 2008). Moreover, both types of team members use face-to-face interactions and computer-mediated communication (CMC) technologies (e.g., online voice chat) to collaborate with each other. However, comparing with geographically distributed team members, the co-located team members have the capabilities and are more likely to engage in face-to-face interactions.

Previous virtual team studies often focused on the impacts of an individual setting of the geographic dispersion alone –co-located or distributed – on team performance (Hambley et al. 2007; O'Leary and Cummings 2007; Purvanova and Bono 2009; Staples and Zhao 2006), or compare the effects of virtual collaboration with face-to-face collaboration in these two settings separately. They overlooked the situation where the situation when both settings (and collaboration styles) coexist in the same hybrid team. O’Leary and Cummings (2007) argued that team members with different geographic dispersions have different communication patterns and thereby result in different team performance. They suggested that there are two important categories of geographic dispersions: spatial and temporal dispersions. A large spatial dispersion is more likely to decrease the likelihood of face-to-face communications among team members (Te"eni 2001). A large degree of temporal dispersion will reduce the potential for synchronous communications among team members (Malone and Crowston 1994). One of the major reasons we want to study the impacts of geographic dispersions on team performance is that both the spatial and temporal dispersions can be regarded as major actionable factors for project leaders to organize the globally distributed team (Colazo and Fang 2010). However, there is no rigorous empirical support for this practice. Our study delves into the issue, which is not only academically interesting but also practically substantive.

2.3.3 Geographically Co-Located Setting (Face-to-Face Interactions among OSS Developers)

As mentioned above, a key assumption of this research is that geographically co-located are more likely to have face-to-face interactions which may facilitate the effects of social influence among OSS developers. The majority of existing studies on face-to-face social interactions focused on teams in traditional organizations in which most members are co-located in a workplace. This is mainly because geographically co-located team members work in close physical proximity and are more likely to have frequent face-to-face interactions. Crowston et al. (2005) suggested that face-to-face collaborations also existed in many OSS project teams which many studies consider as a largely distributed, virtual collaboration oriented environment. Crowston et al. (2007) also explored the the role of face-to-face meetings as a form of professional communication in the life of technology-supported self-organizing distributed (or virtual) teams. This means both settings may co-exist in real world OSS project teams and significantly affect their performance and success at the same time. However their relative impacts are rarely investigated in such situations.

It was suggested that face-to-face collaboration in general allow people observing each other's physical and emotional cues which can improve team members' innovation abilities and contribution motivations (Hart 2001). It was also found that the social and interpersonal cues that provide the basis for social influence are lacking in geographically distributed teams (Branson et al. 2008). Moreover, Staples (2006) found that geographically co-located team members share higher satisfaction of their teams.

Moreover, face-to-face collaborations can provide an ideal environment for offline social influence among team members, and thereby improve team performance. Olson et al. (2002) found that "when people have questions, often the person who could answer it (i.e., a fellow worker who had more experience or expertise on a topic) was at hand". Warkentin et al. (1997) investigated 11 geographically co-located teams and discovered that they have high degree of cohesion. In addition, it was found that task conflicts among team members rarely existed in geographically co-located teams (Staples and Zhao 2006; Wakefield et al. 2008).

2.4 The Impacts of Online Social Influence on OSS Project Participation

The second study of the thesis is about modeling and investigating social influence through explicit social relationships. This approach is rooted in two streams of social influence studies: word-of-mouth (WOM) and observation learning (OL).

Two major types of social influence mechanisms were extensively studied in marketing and economic literature. First, Arndt (1967) defined the mechanism that consumers' product adoption is influenced by others' opinions and experiences as word of mouth (WOM). Nowadays, such opinions are often in the form of online reviews or social media communications. Second, Bandura (1971) and Barbagallo et al. (2008) defined the mechanism that individuals may observe and be influenced by others' actions without knowing the motives and reasons behind such actions as observational learning (OL). In this study, we also adopt this framework to study how these mechanisms affect OSS developers' project participations.

2.4.1 The Impacts of Word of Mouth (WOM) on OSS Project Participation

Word of mouth studies mainly focused on its impacts on consumer behaviors and product sales (Awad and Ragowsky 2008; Cheung and Thadani 2012; Rui et al. 2013). These studies indicated that WOM valence (positive or negative) can change consumers' evaluation of the products (Chevalier and Mayzlin 2006; Mizerski 1982), while WOM volume may help facilitate better consumer awareness and increase the number of informed consumers. Moreover, the emergence of the Internet services like online reviews for “publicizing feedback and recommendations on products” has attracted many researchers to study WOM in the digital age (Chen and Xie 2008; Dellarocas 2003; Duan et al. 2008). For instance, Clemens et al. (2006) conducted a survey of online reviews from craft beer industry and found that products with high valence are likely to be bought again. Cheung et al. (2014) found that an increase in the volume of online product ratings can improve sales.

However, the impacts of WOM on OSS participations have not been well studied. Krishnamurthy (2003) suggested that in general there is a lack of resource for marketing OSS projects through traditional media. Then Bagozzi and Dholakia (2006) and Barbagallo et al. (2008) briefly discussed that WOM can be useful in advertising OSS projects and building awareness among developers. More recently, Santos et al. (2013) pointed out that WOM has great potential in influencing developers’ participation behaviors. However, they all did not empirically investigate the impacts of WOM.

In the context of our study, microblogging service in the Open Hub community enables developers to disseminate their opinions, recommendations, and activities of OSS projects among developers. Jansen et al. (2009) and Hennig-Thusrau et al. (2015) suggested that microblogging offers a novel electronic channel of WOM. While earlier microblogging research mainly focuses on individuals’ motivations to post (Davidson and Vaast 2009; Java et al. 2007; Zhao and Rosson 2009). Existing studies (Dabbish et al. 2012; Seebach et al. 2011; Tsay et al. 2012) have indicated that microblogging can enhance transparency and collaboration for software developers. We conjecture that using microblogging to publish developers' positive experiences, opinions, or participation activities can raise project awareness and in turn attract more participation.

2.4.2 The Impacts of Observational Learning (OL) on OSS Project Participation

Observational learning research explores its impacts on consumer product adoption behaviors Bikhchandani et al. (1992). Their theoretical explanation suggests that OL information contains signals expressed by others' adoptions but not the reasons behind such actions. When there is limited product information, the publicly observed other consumers' adoptions by an individual outweighs her own private information in her adoption decision. As more and more consumers follow their predecessors' adoptions, an information cascade and behavior "herding" occur among people (Banerjee 1992).

The impacts of OL are amplified in online environment, as individual's online activities are becoming increasingly transparent (Cheung et al. 2015; Dellarocas et al. 2010; Ye et al. 2013; Zhou et al. 2013). For instance, Burke et al. (2009) found that social networking site users who see their friends' contributions are motivated to share more content. Dellarocas (2010) and Cheung et al. (2015) found OL affects people's information contributions in online communities.

However, the impacts of OL on OSS participation are not investigated while developers' participation becomes increasingly visible on various platforms. One related finding in Hahn et al. (2008) is that OSS developers tend to participate in projects with members they have worked with in other projects. This may indicate actions of a developer's acquaintances may affect her future participation choices.

Comparing with OL, the information conveyed through WOM are more of subjective (personal) opinions or evaluations. The publicly observed others' actions information is often in the form of objective statistics (e.g., sales). OSS developers' participation is a process which developers become more engaged and thus more familiar with the project and its members. We conjectured that developers' reliance on social influence may change in this process as they can better evaluate the project and its members' words due to such familiarity. Moreover, the difference in the objectivity of the information conveyed by OL and WOM may cause different changes in developers' reliance on social influence. The details of those changes will be examined in our study. In order to do that, we first review the literature about initial and sustained OSS participation, their differences, and how social influence may affect them differently.

2.4.3 Initial and Sustained OSS Project Participation

Existing OSS research on developers' participation behaviors mainly focused on the motivations of developers' initial participations (i.e., initial reasons for joining the projects) (Ghosh 2005; Hann et al. 2004; Hertel et al. 2003; Lakhani and Wolf 2005; Subramanyam and Xia 2008). Comparing with initial participations, there are very few studies that just have begun to explore what mechanisms may sustain long-term voluntary developers' project participations (Fang and Neufeld 2009). Among these studies, Shah (2006) found that long-term participants enjoyed programming and interacting with other developers. This empirical finding suggest that social influence among project members like WOM and OL may play an important role in sustained participations. They also found that initial participations were predominately driven by immediate software use value. Such differences suggest OSS developers may initially join in a project with some short-term needs, but such needs may transform to long-term mechanisms like enjoyment over time. It also implies social influence may have differential impacts on initial and sustained OSS participation.

In a similar vein, Bagozzi and Dholakia (2006) found that sustained participation is associated with developers' senses of identification. Such senses are often strengthened by complex social interactions among project members. Engaged project members view their contribution as "enjoyable joint activities to be done" with their peers. Fang et al. (2009) also found that long-term contributors are influenced by their social interactions with the project community. Von Hippel and von Krogh et al. (2003) found that the momentum for developers' sustained participation is largely due to their social interactions with other project members. However, all these studies did not examined the impacts of social influence that are embedded on such interactions, which has contributed to strengthened sense of identity and sustained participation.

To summarize, prior OSS participation motivation research focused on the initial participation. The social influence perspective is largely ignored. Meanwhile, the few sustained participation studies have found that social interactions among project members may strengthen developers' senses of identification and long-term enjoyment, thereby contributing to sustained participations. However, all those studies did not investigate the impacts of social influence on OSS project participation. This thesis aims to fill this gap by investigating the impacts of online social influence on developers' initial and sustained project participation behaviors.

Chapter 3: Problem Statement

The general research goal of the thesis is to examine the impacts of both offline (geographic dispersion) and online (WOM and OL) social influence on the OSS project success. We use data from a large online OSS development community to model and analyze social influence through both geographic dispersion and explicit social relationships. In this section, we present the research data and the research questions of the two studies in the thesis separately.

3.1 Research Testbed (Data)

The data analyzed in the thesis is mainly collected from a large online open source software (OSS) community – The Black Duck Open Hub (Open Hub)[1]. To do so, I developed a set of Java programs to automatically query and retrieve data through the API provided by the Open Hub. Since all retrieved data items are in XML format, a parser program was developed to parse them into a database. Open Hub offers analytics and search service for discovering, evaluating, tracking, and comparing open source code, developers, and projects. OSS developers can join or add new projects. Open Hub also encourages OSS developers to register, edit and discuss their projects in its online social-networking community. Moreover, each community member can rate the quality of an OSS project/product she used through a score (ranging from 1 to 5, where 5 is the best score).

It also provides two types useful information which is not available at other major OSS portals. These two types of information provide us the opportunities to study offline and online social influence. First, Open Hub allows developers to register their addresses in their database and parse them to coordinates (with longitude and latitude information). Therefore, such location information along with the project rating provides us a unique opportunity to investigate the

[1] https://www.openhub.net/

relationship between the average team members' distance (representing different geographic dispersion patterns) and team performance (project rating), as required by the research questions of the study I.

Second, Open Hub provides a microblogging service on its platform from 2008 to 2012. It allows developers to publish information about their opinions, recommendations, and project participation activities through profile web pages of projects and followers, thereby may influence others' participations through word of mouth effects. In the profile page of each project, Open Hub displays the project development activity summary in the project profile webpage, such as number of commits (a commit is a one-time developer's contribution to the source code of an OSS project), number of developers. Therefore, developers can also be influenced through observational learning effects.

3.2 Research Questions

3.2.1 Study I: The Impacts of Geographic Dispersion on OSS Team Performance

Prior virtual team study has specified geographic dispersion as a multi-dimensional construct, spatial, temporal, and configurational (O'Leary and Cummings 2007). Research in virtual teams also suggested that the probability of face-to-face communications drops fast as the degree of geographic dispersion increases (Allen 1977; Van den Bulte and Moenaert 1998). However, most studies mainly investigated the associations between geographic dispersions and team performance in individual settings of geographic dispersion. In our study, we adopt the instrumental variable estimation (IV) method to investigate the causal effects of geographic dispersion on team performance (project success) in a large scale, real world OSS development community where both geographic dispersion settings co-exist in many projects.

In addition, virtual team and organization research argued that the spatial dispersion reflected by the geographic distance among team members may bring reduction in spontaneous communication, because it will decrease the likelihood of face-to-face interaction for them (Allen 1977; Kraut and Streeter 1995; O'Leary and Cummings 2007; Te"eni 2001). Temporal dispersion which is reflected by the time differences among team members is more likely to reduce the possibility of the real time problem solving, since it may decrease the likelihood of synchronous interactions for them (Grinter et al. 1999; Herbsleb et al. 2000; Malone and Crowston 1994; O'Leary and Cummings 2007).

Media richness theory (MRT) (Daft and Lengel 1986; Dennis et al. 2008) argued that face-to-face communication being the richest, while other media capable of sending fewer cues (e.g., vocal inflection, gestures) or providing slower feedback (e.g., memos, voice-mail, or e-mail) are “leaner”, therefore face-to-face communications can improve team performance. In our study, we adopt the MRT to support our research. Currently, most of the empirical MRT research examine the managers’ perceptions of media fit by surveying their media choices, but largely ignore the performance of different media use.

This thesis aims to fill the gap by empirically examining the impacts of face-to-face interactions reflected by the geographic dispersion of team members within a virtual team. Moreover, we adopt the IV estimation method to build a causal link between the geographic dispersion and team performance. Furthermore, Open Hub data set provides the geographic information e.g., latitude and longitude coordinates of OSS developers within the project. The feature of the data provided by the Open Hub along with the rigorous econometric method enable us to measure the degree of the geographic dispersions among OSS developers within the project, thus to verify the hypothesis. To summarize, the two main research questions to be addressed by the study I are as follows:

Research Question 1.1 (RQ 1.1): What are the impacts of geographic dispersion of OSS team members on their team performance (project success)?

Moreover, existing research has not distinguished the impacts between spatial (face-to-face collaboration) and temporal (convenience factors) geographic dispersion on OSS team performance. Therefore, the second research question is:

Research Question 1.2 (RQ 1.2): Do spatial geographic dispersion and temporal dispersion have different effects on OSS team performance (project success)?

3.2.2 Study II: The Impacts of the Online Social Influence on OSS Project Participation

Social coding websites, such as GitHub, Open Hub provide social media services for OSS developers to collaborate with each other (Yoshikawa et al. 2014). The social influence enabled by the social coding website can facilitate the interaction and collaboration among OSS developers and help them to create innovative software projects. Therefore, we argued that online social influence (WOM and OL) may have differential impacts on initial and sustained OSS participation, mainly due to two reasons. First, the objectivity of the information conveyed through the two online social influence mechanisms are rather different. Second, developers' knowledge level of the underlying project and its members may increase from initial to sustained participation (stage). They may become more familiar with the project and its members and more capable to evaluate these members' subjective opinions (WOM), thereby can better decide whether to continue to contribute to this project (i.e., sustained participation). However, this needs to be empirically examined. Therefore, we propose the following research questions:

Research Question 2.1 (RQ 2.1): What are the impacts of social influence (WOM and OL) on OSS project initial participation?

Research Question 2.2 (RQ 2.2): What are the impacts of social influence (WOM and OL) on OSS project sustained participation?

Chapter 4: Research Design

4.1 Overall Research Design of the Thesis

Figure 2 illustrates the overall research design of the thesis. There are three main steps included by the research design of the PhD thesis. Firstly, a set of Java and R programs are developed to automatically query and retrieve data through the API provided by Open Hub. Since all retrieved data items stored in Open Hub platform are represented in XML format, a parser program was developed to parse them and store into a database. Such information includes OSS developers' project participation and contributions, their positive evaluation choices, location, nationality, programming language preferences, development activities, and project statistics. Moreover, it also keeps track of the changes in the source code of each listed OSS project from the version control systems and calculated software metrics such as the total number of changes (i.e. commits) at different time periods.

In the second step, we model the impacts of social influence on the OSS project success. We mainly investigate two types of social influence. Based on previous reviewed literatures, the offline social influence is mediated by the face-to-face interactions among OSS developers. The likelihood of face-to-face interactions is indicated by the degree of geographic dispersion among OSS developers. We argue that higher likelihood of face-to-face interactions among OSS developers will facilitate the offline social influence among them thereby benefits the OSS project success. As the social coding website along with social networking-like functions spawns, the online word of mouth (WOM) and observational learning (OL) may influence individuals' behaviors as well. In this thesis, we model the impacts of WOM and OL on OSS developers' project participation behaviors. OSS developers' project participation behaviors represent the engagement of OSS developers which is also regarded as the key indicator of the project success (Daniel et al. 2013).

In the third step, we use rigorous econometric methods to analyze the data obtained. The instrumental variable estimation and panel regression model are used in the two studies to provide causal interpretations of the results. Then we discuss the research implications of the thesis and provide practical insights for OSS stakeholders. In the next two sections, we will present the research design of the two studies respectively.

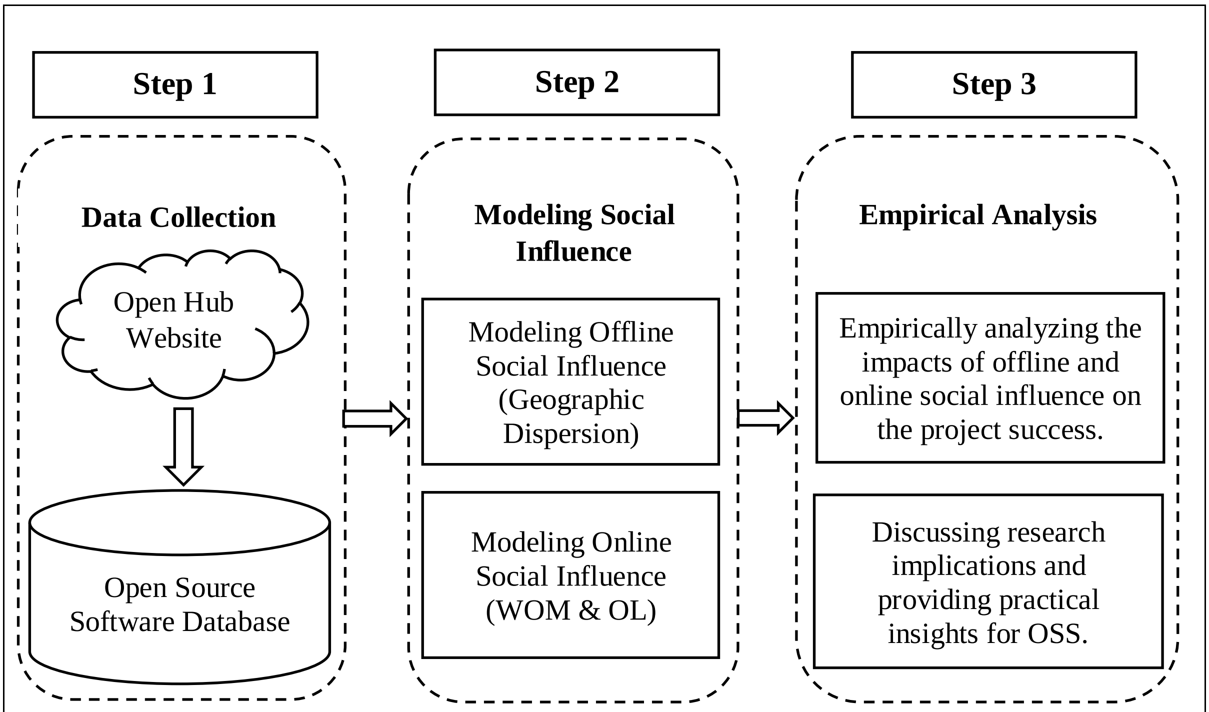

Figure 2: Overall Research Design of the Thesis

4.2 Research Design of the Impacts of Geographic Dispersion on OSS Team Performance (Study I)

In order to address research questions 1.1 and 1.2, we develop a research model as shown in Figure 3. It generally suggests that lower spatial dispersion (likely more face-to-face interactions that facilitate social influence) and lower temporal dispersion can lead to better Project success (team performance). The effects of these two mechanisms in this model need to be distinguished, along with various project related control variables we extracted from relevant studies as shown in Figure 3. The solid arrows in the figure represent the causal relationship between the spatial dispersion among OSS developers and the OSS project success, it indicates that the smaller the degree of the geographic distance a project has, the higher project success it will get. Based on previous OSS literatures (Crowston et al. 2003; Crowston et al. 2006), the project level factors which also affect the OSS project success, such as project age, size, team diversity, etc. are included in this study.

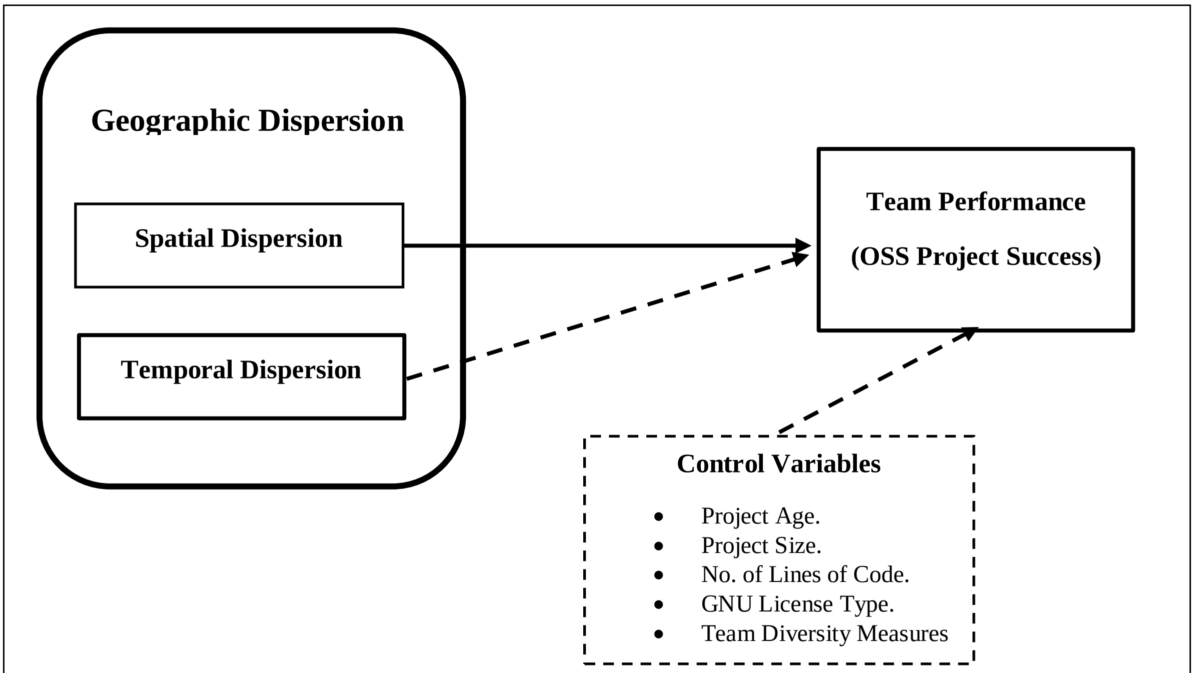

Figure 3: Research Framework of the Study I

4.2.1 Measurements of the Study I

Dependent Variable: OSS Project Success (Team Performance)

We adopt the average users' rating score of an OSS project in Open Hub as the measurement for the success of that project. Any user registered in Open Hub can rate an OSS project based on her overall satisfaction with a rating score ranging from 1.0 to 5.0 (e.g., 3.6), in which 5.0 is the best rating. The average rating score is used as the dependent variable and denoted as RAT. Table 1 shows the detailed definitions of variables used in our empirical analysis.

Independent Variables

The developer's location information provided by the Open Hub website includes three perspectives: 1) an optional text description of this contributor's claimed location; 2) the country code – a string (e.g., US) representing his country; and 3) a pair of floating point values representing the latitude and longitude of his exact (home or work) location. The location data has been validated by either the Google geocoder web service.

Most studies measure the spatial dispersion by a dichotomous variable to distinguish team processes and outcomes between collocated and distributed teams (Ehrlich et al. 2008; Herbsleb and Mockus 2003). Some studies also consider the number of sites as the degree of spatial dispersion (Nguyen et al. 2008; Ramasubbu et al. 2011). While others examined whether team members were located in the same room or different building, city, or country (Chudoba et al. 2005; Espinosa and Carmel 2003).

In our study, we use the latitude and longitude values of members in an OSS project for calculating the average geographic distance (spatial dispersion) and average time zone difference (temporal dispersion) among these members, based on the geographic dispersion measures constructed by O'Leary and Cummings (2007).

Control Variables

Moreover, other project level factors that were explored in previous research are also included in our study as control variables. These factors include types of software, project age, project size (Crowston and Scozzi 2002), GNU license type (Grewal et al. 2006). The development of social coding websites such as GitHub, Open Hub, etc. has enabled a virtual community which facilitates the collaborations among OSS developers. The ability to leverage diverse knowledge and experience is critical to the OSS communities. Daniel et al. (2013) adopted a diversity lens to study the success of open source software (OSS) projects. Based on previous literatures

(Harrison and Klein 2007; Van Knippenberg and van Ginkel 2010), they proposed that three types of diversity (disparity, separation, and variety) are important to OSS project success. In our study, we operationalized team diversity related control variables as well. The definition and description of variables used in the study I are presented in the following table 1.

Table 1: Variable Definition and Description (Study I)

Variable Definition	Description
Dependent Variable	
Average Rating Score **(RAT)**	A floating point value from 1.0 to 5.0 represents the average value of all ratings by the Open Hub community members. 1.0 is the lowest possible rating, while 5.0 is the highest possible rating.
Geographic Dispersion Measures	
Spatial Dispersion	$\frac{\sum_{i=1}^{K}\sum_{j=i}^{K} \mathrm{Site\_Dist}(i,j) * n_i * n_j}{(N^2 - N)/2}$ 1): Site_Dist (i,j) measures the geographic distance between site i and j with the great circle distance in kilometres 2): n_i , $n_j$is the number of developers in the i^{th} and j^{th}site. 3): K is the total number of sites. 4): N is the total number of developers across all sites included in the project.
Temporal Dispersion	$\frac{\sum_{i=1}^{K}\sum_{j=i}^{K} \mathrm{Time\_Zone}(i,j) * n_i * n_j}{(N^2 - N)/2}$ 1): $\mathrm{Time\_Zone}(i,j)$ measures the number of time zones between site i and j. 2): n_i , $n_j$is the number of developers in the i^{th} and j^{th}site. 3): K is the total number of sites included in the project. 4): N is the total number of developers across all sites included in the project.
Control Variables	
Project Age	The project age of an OSS project refers to how many months this project has existed in the Open Hub community.
Project Size	The project size refers to the total number of developers of an OSS project.
No. of Lines of Code	The number of lines of source code of an OSS project. The number of blank and comments lines are excluded.
GNU License Type	We control for whether or not the project uses the GNU GPL, as prior study (Stewart et al. 2006) have shown that license choice affects project success. GNU GPL is regarded as the restrictive license for OSS projects. Dummy variable with 1 indicates project with GNU GPL otherwise 0.

Disparity Diversity	(1): Developers' contribution is represented by the number of commits of each developer within an OSS project. (2): The disparity diversity (contribution) is calculated by the coefficient of variation of the commits of developers within the project. It is specified as follows: $\frac{\text{Standard Deiviation}(\text{com}_1, \text{com}_2, \ldots \text{com}_N)}{\text{Mean}(\text{com}_1, \text{com}_2, \ldots \text{com}_N)}$ com_i represents the number of commits of the developer i to the OSS project. N is the total number of developers within the OSS project.
Separation Diversity	(1): The imbalance indicates the locations with uneven distribution of developers within an OSS project. (2): It is calculated as follows: $\text{Imbalance} = \text{Standard Deviation}(n_{i,} n_{j,} \ldots n_k)/N$ k is the total number of sites represented in the team. n_i is the number of team members in the ith site. n_j is the number of team members in the jth site. N is the total number of team members across all sites.
Variety Diversity	For each project, the variety diversity is measured by the Blau index based on the most experienced programing language of OSS developers. The variable is operationalized by the Blau index. Higher value indicates that OSS project will be more likely to have developers with broad programming skills. A broad repertoire of programming language available in an OSS community will benefit it by creating a more innovative environment thereby enhance the OSS project success.

4.2.2 Empirical Models and Instrumental Variable (IV) Estimation

In the study I, we adopt a linear regression with instrumental variable estimation method to investigate the causal relationship between geographic dispersion and project success. We use the two-stage least squares (2SLS) procedure for the instrumental variable estimation.

In order to investigate the causal effects of geographic dispersion on project success, we develop an instrumental variable to examine the possible endogeneity our main independent variable - average geographic distance – may bring.

A valid instrumental variable must satisfy the following condition: it should be correlated with the independent variable but uncorrelated with the dependent variable. The instrumental variable applied in our study for the geographic distance of the current project p is the sum of average geographic distance of projects these developers have participated before. We have empirically tested and found that the average geographic distance among developers within a project (independent variable) is correlated with the instrumental variable. Both the independent variables and the instrumental variable represent the degree of the geographic dispersions of the projects these developers choose to participate. However, the instrumental variable represents the degree of geographic dispersion of other projects except the current project p, it is uncorrelated with the rating of the current project (dependent variable). Therefore, the instrumental variable used in our study satisfy the condition for a valid instrumental variable.

The instrumental variable $\mathrm{ivdist_p}$ in our study is calculated by the following formulas:

$$\mathrm{ivdist_p} = \frac{\sum_{i=1}^{N} \mathrm{Dist_i}}{N} \quad (1),$$

$$\mathrm{Dist_{i=}} \frac{\sum_{j=1}^{M} \mathrm{GeoDist_j}}{M} \quad (2),$$

Where N is the total number of developers of the project p, and $Dist_i$ is the average value of average geographic distance for the outside projects the developer i has participated before. M is the total number of projects that the developer i has participated before except the project p. $\mathrm{GeoDist_j}$ is the average geographic distance of the developer i's outside projects j.

4.3 Research Design of the Impacts of Online Social Influence on OSS Project Participation (Study II)

In this section, we present the measurements of OSS project initial and sustained participations, word of mouth (WOM), observational learning (OL), as well as the empirical model which is used to the study II. All the variables in the study II are summarized in the table 2.

4.3.1 Measurements of the Study II

To study the impacts of social influence (WOM and OL) on OSS developers' initial project participation, we use the number of developers who participate in a specific project for the first time in month t ($\text{Monthly\_New\_Participation}_{it}$) as the dependent variable. On the other hand, we use the number of new commits to that project in month t ($\text{Monthly\_New\_Commits}_{it}$) as the dependent variable for measuring the level of developers' sustained participation.

Independent Variables (WOM and OL)

In our empirical analysis, we use the volume of the microblogging messages with the project tag published by existing project members in month t-1 ($\text{Microblogging\_Tags}_{i,t\text{-}1}$) as the independent variable for the WOM mechanism. Previous social influence studies usually model both the volume and valence of the WOM messages. We have browsed the Open Hub microblogging messages and found there is little negative content. Most messages are positive opinions or about project progress. Therefore, our model only includes volume of microblogging messages.

In the OL process, the actions of prospective participants should be consistent with the actions of existing members they have observed. Thus for the dependent variable initial participation, we use the total number of existing developers in the underlying project in month t-1 ($\text{Cumu\_Developers}_{i,t\text{-}1}$) as the independent variable. Because this number shown in the project profile web page indicates the cumulative number of initial participations since the start of the project. Similarly, we use the cumulative number of commits ($\text{Cumu\_Commits}_{i,t\text{-}1}$) as the independent variable for sustain participation since it measures the level of cumulative developer contributions in a project.

Control Variables

Cumulative number of code lines ($\text{Cumu\_Codelines}_{i,t-1}$): This variable is often used to measure the complexity of the underlying software project and the project output. (von Hippel and von Krogh 2003) used this measure as an index. We conjecture that a more successful project tends to attract more participations, and we then include it as our control variable.

Project age ($\text{Project\_Age}_{i,t-1}$): Grewal et al. (2006) indicated that project age signals stage of project life cycle. Hahn et al. (2008) suggested that different developers may prefer joining projects at different stage. Subramaniam et al. (2009) argued project age may be a proxy for other factors affecting project success such as the developers' group experience.

$\text{Monthly\_New\_Participation}_{i,t-1}$ and $\text{Monthly\_New\_Commits}_{i,t-1}$ are lagged dependent variables and used to control for reverse causality issues. Such issues arises when we cannot distinguish if more social influence (WOM and OL) effects that cause more initial and sustain participations, or more participations have generated more WOM and OL.

Average experience of project members ($\text{Project\_Members\_Exp}_{i,t-1}$) is used in previous study (Roberts et al. 2006) as a proxy for developers' characteristics like knowledge and skills that are difficult to measure and may affect the project performance. We also adopt this measure as a control variable for assessing two of the commonly used OSS project performance measures – the initial and sustained project participations.

Average reputation score of project members ($\text{Project\_Members\_Rep}_{i,t-1}$): Our previous research (Hu et al. 2012) has found that developers with good reputation score tend to attract more project collaborators since they may want to learn from these reputable OSS developers in terms of programming or project management. We then adopt average number of project members' reputation as our control variable.

Table 2: Summary of Measures (Study II)	
Dependent Variables	
$\text{Monthly\_New\_Commits}_{it}$	The number of new commits made to the project i in month t.
$\text{Monthly\_New\_Participation}_{it}$	The number of new developers who participated in project i in month t.
Independent Variable (WOM)	
$\text{Microblogging\_Tags}_{i,t-1}$	The number of microblog messages which contain a tag that links to project i's name in month t-1.
Independent Variables (OL)	
$\text{Cumu\_Commits}_{i,t-1}$ (Sustain Participation)	The cumulative number of commits the project i has until month t-1.
$\text{Cumu\_Developers}_{i,t-1}$ (Initial Participation)	The cumulative number of developers the project i has until month t-1.
Control Variables	
$\text{Cumu\_Codelines}_{i,t-1}$	The cumulative number of lines of code, excluding comments and blanks of the project i until month t-1.
$\text{Project\_Age}_{i,t-1}$	The number of months i existed until month t-1
$\text{Monthly\_New\_Commits}_{i,t-1}$	One-month lagged variable of the monthly new commits.
$\text{Monthly\_New\_Participation}_{i,t-1}$	One-month lagged variable of the monthly new participation.
Control Variables (Future Study)	
$\text{Project\_Members\_Exp}_{i,t-1}$	The average number of months developers spent on the project i until month t-1.
$\text{Project\_Members\_Rep}_{i,t-1}$	The average number of the project members' reputation scores (Kudo rank) in the project i.

4.3.2 Panel Regression Model

The study II aims to provide a causal interpretation of the observed correlation between the two types of social influence mechanisms and OSS developers' project participation behaviors. We carefully designed our empirical model which leverage the panel structure of our data sample to control for the unobserved heterogeneity in project characteristics and possible endogeneity issues like reverse causality. The dependent variables are the measures of initial and sustained OSS project participation behavior as defined previously, and the WOM and OL measures are used as independent variables. Controlling for project-level unobserved effects is achieved in the panel model by introducing fixed effects. We also control the project-specific fixed effects ρ_i and η_i in the two models to capture the idiosyncratic characteristics associated with each project, such as project license, programming language, manager etc. In addition, the one-month lagged dependent variables are used in our model for the identification of reverse causality issues. In order to decide between fixed or random effects, we ran a Hausman test. The p-value of the Hausman test results is 0.01207. The p-value is significant (p-value<0.05). Then we choose to use fixed effects in our study. The fixed effects capture the time invariant, unobserved heterogeneity of each project. Thus we can control for unobserved differences across different projects.

Chapter 5: Overall Results of the Thesis

In this section, we present the overall results of the cumulative thesis by summarizing the main results of the two studies. The first subsection introduces the overall results of the study I which examines the impacts of offline social influence (geographic dispersion) on OSS project success. The second section discuss the overall results of the study II which investigates the impacts of online social influence: word of mouth and observational learning on initial and sustained participation in OSS project success.

5.1 Results of the Impacts of Geographic Dispersion on OSS Team Performance

5.1.1 Main Results

The study I mainly examines the impacts of geographic dispersion on OSS project success (team performance). Specifically, we use the instrumental variable estimation to provide a causal interpretation between the degree of the geographic dispersion among OSS developers and the average rating score of the project (project success). The main results of the study I indicates that the smaller the average geographic distance a project has, the higher the average rating score it is likely to get. This suggests that, in general, higher likelihood of face-to-face collaboration among the members of an OSS project will increase the chances of its success.

To summarize, the main results of the study I suggest that the spatial dispersion has a negative impact on project success, even after controlling the temporal dispersion. Geographic dispersions can be regarded as major actionable factors for project leaders, especially in the online virtual community.

5.1.2 Follow-up Results

This section presents the follow-up results of the study I. In this study, we first collected OSS projects registered on the Open Hub website. There are almost 636,013 projects collected with the crawling program written by Java. Then we eliminated projects without the average rating score value due to the purpose of our analysis. There are 16,119 projects left. Then we collected the geographic information from Open Hub for OSS developers participated in the 16,119 projects. In addition, we also collected the geographic information from the GitHub website[2] for developers whose location information cannot be obtained from the Open Hub. Finally, there are almost 30,022 OSS developers with geographic information as used in study to calculate the geographic dispersion measures. Since the study needs to measures the average geographic distance among developers of an OSS project. Thus, it requires OSS projects which have at least two developers with geographic information. Therefore, our data sample are filtered and there are 8,001 projects left for the analysis.

Table 3 reports the correlations of the constructed variables. Both the spatial dispersion and temporal dispersion variables are negatively correlated with the dependent variable.

Table 3: Correlation Table of Variables of the Study I

	(1)	**(2)**	**(3)**	**(4)**	**(5)**	**(6)**	**(7)**	**(8)**	**(9)**	**(10)**
(1) RAT	1.00									
(2) Spatial Dispersion	-0.08	1.00								
(3) Temporal Dispersion	-0.06	0.41	1.00							
(4) Project Age	-0.24	0.07	0.08	1.00						
(5) Project Size	0.02	0.06	0.03	0.01	1.00					
(6) No. of Lines of Code	0.01	0.01	0.01	0.03	0.29	1.00				
(7) GNU License Type	-0.03	-0.12	-0.13	0.05	0.01	-0.03	1.00			
(8) Disparity Diversity	0.09	0.15	0.12	0.24	0.27	0.07	0.06	1.00		
(9) Separation Diversity	0.03	-0.27	-0.11	0.08	-0.07	0.03	0.03	0.12	1.00	
(10) Variety Diversity	0.02	0.07	0.04	0.17	0.21	0.09	0.02	0.16	0.15	1.00

In our study, we use the linear regression model with instrumental variable estimation to investigate the causal relationship between geographic dispersion and project success. We use

[2] https://github.com/

the two-stage least squares (2SLS) procedure for the instrumental variable estimation. In order to investigate the causal effects of the spatial dispersion of OSS developers on project success, we develop an instrumental variable to examine the possible endogeneity caused by our main independent variable – spatial dispersion.

Table 4: Follow-up Results of the Study I		
Dependent Variable: Average Rating Score (RAT)		
	Model (1)	Model (2)
Geographic Dispersion Measures		
Spatial Dispersion	-0.0035 *** (0.0014)	-0.0034 *** (0.0015)
Temporal Dispersion	-0.0011 ** (0.0007)	-0.0013 ** (0.0010)
Control Variables		
Project Age	-0.0189 *** (0.0011)	-0.0192 *** (0.0013)
Project Size	0.0041 ** (0.0013)	0.0042 * (0.0016)
No. of Lines of Code	0.0028 (0.0039)	0.0031 (0.0043)
GNU License Type	-0.0042 *** (0.0014)	-0.0041 ** (0.0023)
Disparity Diversity		0.0021 *** (0.0011)
Separation Diversity		0.0017 ** (0.0008)
Variety Diversity		0.0013 * (0.0010)
$\mathbf{R^2}$	0.27	0.46

*Significant at 0.1; **Significant at 0.05; *** Significant at 0.001
Standard errors are reported in parentheses.

Table 4 shows the regression results of the empirical analysis using the instrumental variable estimation method. The definition and description of the variables are presented in the Table 1. We examine the causal impacts of the average geographic distance among OSS team members on the project average rating score. Model 1 consists the independent variable and project success related control variables. Besides the control variables used in the model 1, the model 2 adds the previous constructed diversity measures.

The results of the model 1 suggest that the spatial and temporal dispersions have strong negative effects on the average rating score (project success). It indicates that the smaller the average geographic distance a project has, the higher the average rating score it is likely to get. After controlling for the temporal dispersion, the coefficient of the spatial dispersion is still significant with a negative sign.

Project age has a strong negative effect on the average rating score. It suggests that the younger a project is, the more successful may become. It is consistent with previous OSS research that found the younger projects are more likely to attract developers to join in.

Project size has a strong positive effect on the average rating score. It indicates that the larger number of developer a project has, the higher the average rating score of the project is likely to have. In this model, the number of lines of code has no significant effect on OSS project success.

GNU license type is also controlled in our study, as prior studies have shown that license choice also affect OSS project success. In our study, we control for whether or not the project uses the GNU GPL license. Prior OSS studies (Daniel et al. 2013; Stewart et al. 2006) argued that projects which adopt a non-restrictive license would experience greater success than projects with a restrictive license such as GNU GPL. Our results support this argument as well.

The effects of diversity on OSS project success

Based on the model 1, the second model adds the constructed diversity measures as control variables as well. Our results are consistent with prior research of the effect of diversity on OSS project success (Daniel et al. 2013). The results of the model 2 suggests that each of the three diversity measures has significant effect on the average rating score.

The Effects of the Disparity Diversity on OSS Project Success

Our results show that the developers' contribution based disparity diversity among OSS developers has a strong positive effect on the average rating score. It indicates that higher disparity diversity among developers within an OSS project will be beneficial for the OSS project success. Daniel et al. (2013) suggested that the higher disparity diversity of a typical OSS project, the more likely a hierarchical social structure will be observed in it. They argued that although OSS projects are generally considered as non-hierarchical structure (Gallivan 2001), when the contributions of developers and their resultant reputations exhibit significant variation, the community has a social structure, with those developers who have made more contributions being positioned higher in the community than those who make fewer contribution (Dahlander and O'Mahony 2011; Stewart 2005). The emergent social structure in the OSS community will facilitate decision-making among OSS developers thus enhance project success.

The Effects of the Separation Diversity on OSS Project Success

The measure of the separation diversity used in this study is based on the configuration imbalance measures proposed by O'Leary and Cummings (O'Leary and Cummings 2007; Yoshikawa et al. 2014). The results show that the separation diversity has a significant effect on the average rating score. Higher degree of the separation diversity indicates that OSS developers will be more likely to live in the same site (city). The concentration of developers in terms of location will help OSS teams to reduce the differences of cultural, group cohesiveness, and conflicts among team members which will hamper the team performance (Daniel et al. 2013).

The Effects of the Variety Diversity on OSS Project Success

Previous OSS studies suggested that the variety diversity is expected to positively influence technology context since various knowledge and experience is important for team members to create innovative products thereby enhance project success (Daniel et al. 2013; Joshi and Roh 2009). Our results are consistent with this argument and show that the variety diversity has a significant positive effect on the OSS project success.

5.2 Results of the Impacts of the Online Social Influence on OSS Project Participation

The main results of the study II suggested that the impacts of word of mouth (WOM) on OSS developers' initial and sustained project participation differ from each other. Specifically, a focal OSS project's prospect participants are significantly influenced by its related microblogging messages for their initial participation. As these participants become more familiar with the project and its members after initial participation, such WOM influence disappears. We conjecture it is because that project and member familiarity enable developers to better evaluate the more subjective information conveyed through other members' WOM.

On the other hand, observational learning (OL) effects existed for both OSS developers' initial and sustained participation. This may be because that the OL information is often objective statistics. Its impacts are difficult to change when developers' own familiarity or perception of the underlying project changes in the stage of sustained participation.

Our main results also suggest that older projects are less likely to attract initial and sustained participation from OSS developers. Consistent with prior OSS studies (Fang and Neufeld 2009; Roberts et al. 2006), our results support that the cumulative number of lines of code added to the OSS project can influence the developers' participation behaviors.

Chapter 6: Contributions of the Thesis

The overall research goal of the thesis is to provide explorations and insights for open source software (OSS) by richer understanding the impacts of both offline and online social influence on the team performance, initial and sustained project participation. This thesis contributes original ideas, rigorous methods and practical insights to information systems, computer science as well as virtual team literatures. Below, I summarize the contributions of the two studies in more detail and discuss their implications respectively.

6.1 The Impacts of Geographic Dispersion on OSS Team Performance (Study I)

The study I contributes to OSS literatures by theoretically and empirically enriching the understanding of the relationships between the geographic dispersion of team members and OSS project success when co-located settings are more likely to facilitate face-to-face interactions and exert social influence among OSS developers. The study has empirically validated the questions suggested by Crowston et al.(2007) for more research on the role of face-to-face interactions among OSS developers. Moreover, this research distinguishes the impacts of the spatial and temporal geographic dispersion on (OSS) project success which previous research has failed to do so. The research findings show that the spatial dispersion has a negative impact on the OSS project success, even after controlling the temporal dispersion. To best of our knowledge, this is the first empirical study in the OSS literatures which provides a causal interpretation of the impacts of geographic dispersion of team members on the OSS team performance.

This thesis has provided several research implications for the virtual team research as well. Virtual team literature has views of the impacts of geographic dispersion on the team performance (Cramton 2001; O'Leary and Cummings 2007) but without proper empirical evidence especially for the causal relationship between them. This study provides solid, empirical evidence about the relationship between the geographic dispersion and virtual team performance in the OSS contexts. Geographic dispersion can be regarded as a major actionable factor for project leaders, especially in commercial software development projects. Although previous studies pointed out the role of face-to face collaborations in virtual teams especially in OSS teams. There is no rigorous empirical support for this practice, furthermore the causal effects of geographic dispersion on team performance are rarely explored.

Our empirical findings in the study I, can also provide useful insights for OSS stakeholders to devise practical strategies to improve their team performance and project success. OSS developers are highly relied on computer mediated communication (CMC) for coordination and collaboration. It indicates that OSS developers usually adopt less F2F communications. However, our study suggests that face-to-face communication among them can improve the team performance. Based on the above findings, the study I can offer a set of suggestions for both individual OSS developers and project managers about managing project members, building trusting environment and facilitate good influence, and thereby improving team performance (project success). For individual developers, our suggestions focus on increasing the visibility of their locations on the Open Hub community, as well as their previous success, programming skills, and positive evaluations (kudos), aiming to increase their chances to be discovered by more local developers and subsequently build trust and facilitate efficient collaborations.

6.2 The Impacts of Online Social Influence on OSS Project Participation (Study II)

To our best of knowledge, the study II is among the first one to adopt a social influence perspective to study OSS developers' participation behaviors. It can enrich the theoretical understanding of the impacts of social influence on individual behaviors that are not mainly geared towards to maximize economic utility. Second, this study has investigated the possible differential impacts of the WOM and OL on OSS participation behaviors. Third, this research adopted rigorous econometric methods to provide a causal interpretation for the research findings. The findings empirically extends and complements the work of Von Krogh et al. (2012) and Fang and Neufeld (2009) by understanding the impacts of online social influence on the initial and sustained OSS project participation.

The study II also provides important practical implications for OSS stakeholders to leverage social influence to manage OSS participation and devise strategies accordingly. Major OSS platforms like Open Hub and GitHub has adopted various social networking functions like microblogging and reputation systems for a long time. OSS project managers can leverage resources to encourage more positive microblogging (WOM) messages and publish detailed OSS participation activities (OL) in their project profile pages. They may encourage OSS celebrities to publish more positive OSS projects relevant messages on the website to attract more initial participation. To leverage the impacts of OL, social coding websites should make the OL information more salient with better user interface design. However, when they aim to attract more sustained participation, it may be better to shift more resources from WOM to OL based methods.

Chapter 7: Limitations and Future Research

The two articles of the cumulative thesis are subjected to limitations therefore we discuss the limitations of the studies outlined before and provide the future research of the thesis.

We discuss the limitations of the study I and provide directions for future research. First, our empirical results about the relationship between geographic dispersion and project success is consistent with our hypotheses. While this is reasonable, the suggestions we made to both individual developers and project managers for increasing face-to-face interaction opportunities and its impacts on project success is only conceptually analyzed and have not been empirically tested. Future research using experiment or survey methodology should be conducted to validate these suggestions and further improve our understanding between geographic dispersion and OSS project success. Another limitation is that several developer characteristics such as gender and race are not available in the Open Hub website and thus not included in our study. Thus, an important extension of this paper would be studying how different developers' geographic dispersion affect project success when such personal data becomes available.

To summarize, for the study I, our future work may consist of three research directions, including 1) empirically validating the effectiveness of our suggestions for inducing face-to-face interactions and building trustful relationships among local developers, 2) investigating how different developers' geographic dispersion affect project success, and 3) studying the peer effects among co-located developers of the project and their impacts on project participations and contributions.

For the study II we still have a lot of chances to fully finish data processing of all proposed variables, and the empirical model used in our preliminary analysis of the study II still incomplete. We may miss control variables that can be major drivers of initial and sustained contributions. This could affect our current preliminary results.

For the future work of the study II, we first would like to improve our model by adding more control variables based on the OSS and social influence literature and our empirical setting. For instance, as shown in the last two rows of table 4.2, we will examine if having more experienced or highly recognized developers can help an OSS project to attract more initial and sustained participation, and how they may affect the impacts of WOM and OL. Second, we would like to examine the interaction effects of WOM and OL to see if they complement or compete with each other in terms of improving OSS participation. Third, we would like to conduct sub group analysis to find out if project characteristics such as project age, size, and structure, may moderate the impacts of social influence.

Since this thesis mainly investigates the impacts of offline and online social influence on OSS project success respectively. It is interesting for us to examine the interaction effects between offline and online social influence on OSS project success for future research of the thesis. For instance, if the two types of social influence are substitutes then it would be important to investigate which type is more important for different types of OSS project success. If the two types of social influence are complements then strategies to amplify the effect of one type of influence with the other might be helpful for OSS project management. Moreover, in order to make the contributions of our findings more solid and creditable, we will also conduct our study with other methodologies or other open source software data sets.

Chapter 8: Conclusions

This thesis aims to investigate the impacts of social influence on the open source software (OSS) project success. In order to address the research questions of the thesis, I reviewed the literatures from OSS, virtual team and organization studies. Then, I collected scalable amount of data from an online OSS website Open Hub. Thus, I conducted quantitative data analysis. Moreover, I adopted rigorous econometric methods to provide a causal interpretation for our main findings. Specifically, the thesis conducted two studies aiming to understand the impacts of offline and online social influence on OSS project success.

The Study I mainly investigates the impacts of geographic dispersion on the OSS project success. The main findings of the study I show that the average geographic distance (representing the likelihood of face-to-face interactions) negatively affects the team performance (project success), even controlling for the temporal geographic dispersion. It indicates that higher likelihood of face-to-face interactions among OSS developers are more likely to improve the OSS project success. To summarize, the study I contributes to the OSS and virtual team literatures by enriching the understanding of the theoretical and empirical relationship between the geographic dispersions of team members and (OSS) team performance (project success) when co-located and distributed dispersion co-exist in same teams.

Study II focuses on understanding the impacts of online social influence on the OSS project success. More specifically, it examines the effects of online word of mouth and observational learning mechanisms on OSS developers' initial and sustained project participation behaviors. The findings of the study suggest that the online social influence has significant but rather different impacts on initial and sustained OSS participation. Specifically, the impacts of WOM on developers' sustained participation faded away after initial participation as they can better evaluate the underlying project and its members' opinion. The impacts of OL exists for both initial and sustained participation.

The thesis provides empirical evidence and theoretical insights for OSS project success literatures which is not only academically interesting but also practically substantive.

This page is intentionally left blank.

Part 2: Study I: The Impacts of Geographic Dispersion on OSS Team Performance

The Impacts of Geographic Dispersion on OSS Project Success:

Face-to-face vs. Virtual Collaboration

Daning Hu
University of Zurich
Zurich, Switzerland
hdaning@ifi.uzh.ch

Xiao Li
University of Zurich
Zurich, Switzerland
davidlixiao@ifi.uzh.ch

Xiaoquan (Michael) Zhang
Hong Kong University of Science and Technology
Clear Water Bay, Kowloon, Hong Kong
zhang@ust.hk

This study is based on the submission to: Proceedings of the 37th International Conference on Information Systems (ICIS 2016), Dublin, Ireland

Abstract

The success of open source software project teams largely relies on team members' efficient and productive collaborations. In this paper, we investigated the impacts of OSS project members' geographic dispersions on their team performance (OSS project success). Existing empirical studies focused on analyzing the impacts of either virtual or face-to-face collaborations alone, but rarely studied the situation when both settings coexist in a heavily technology-dependent context like OSS development. Moreover, the impacts of spatial and temporal (time zone convenience) geographic distances were not clearly distinguished before. To address these, we use the instrument variable estimation to analyze data from a real-world online OSS community. The results show that the geographic distances among team members negatively affect project success, even after controlling temporal dispersion and other related factors. Our results can provide insights for OSS project managers to help devise strategies and policies to improve team performance and thus project success.

Keywords: Geographic dispersion, Face-to-face interaction, Virtual collaboration, Open Source Software, Instrumental variables

Introduction

The success of an open source software (OSS) project largely relies on the efficient and productive collaborations among its team members. In many OSS projects, team members who are geographically distributed usually rely on various computer mediated communication (CMC) technologies to collaborate, while some geographically co-located members often prefer face-to-face interactions for collaborations. Previous research on team performance focused on the impacts of either virtual (geographically distributed) or face-to-face (co-located) collaboration alone on team performance (O'Leary and Cummings 2007) but failed empirically examine their relative impacts when both settings coexist in the same team, especially in a technology dependent environment like OSS development. It is important for OSS stakeholders to understand how they perform under different conditions and devise effective strategies to improve team performance (project success).

Moreover, geographic co-located team members are also in the same or adjacent time zones. This has given them many advantages in collaborations, such as communicating with each other around the same work time (9 to 5 o’clock). Among geographic co-located members, the impacts of such (time zone based) convenience factors on team performance are difficult to be distinguished from face-to-face interactions if such interactions cannot be directly observed. These two perspectives of geographic dispersion - spatial and temporal (time zone) - may lead to very different strategy to improve OSS project success.

To address these two problems, we empirically analyzed the impacts of the geographic dispersion on team performance using real-world data from a large scale OSS community which consists of 672,141 projects and 661,804 developers. We use the average distance among the team members within the OSS project to measure its spatial geographic dispersion.

The project success is represented by the average rating score from its users. We adopted the instrumental variable method to address the endogeneity problem in our empirical analysis and identify the causal relationship between spatial geographic dispersion and OSS project success (measured by project rating scores). Moreover, we distinguish the impacts of the spatial and temporal geographic dispersion by controlling the time zone differences among team members. Our results show that the closer OSS team members are, the higher average rating score their project is, even after controlling for their time zone differences and various other contextual factors (e.g., project age and size). This indicates the likelihood of face-to-face interactions (represented by spatial geographic distances among OSS developers) can significantly increase the likelihood of project success.

We claim three contributions for this research. First, our work contributed to OSS and collaboration literature by enriching the understanding of causal relationship between the team members' geographic dispersion on their collaboration performance in heavily technology-dependent environments like OSS development. Second, we are among the first to distinguish the impacts of the spatial and temporal geographic dispersion on (OSS) team performance. Third, our empirical findings on the moderating roles of various project related factors (e.g., project size) may help team managers devise strategies and policies to improve collaboration efficiency and productivity of their teams and thereby improve project success.

The remainder of this paper is structured as follows. In the next section, we review the studies that are relevant to this research. We then introduce our unique empirical data set and present the overall research design. The fourth section will show our empirical study that aims to discover the impacts of various moderators of causal peer influence. At last, we will discuss our ongoing work in developing a simulation approach and our own seeding strategy that incorporates the empirical insights learned from the above analyses, and compare the effectiveness of this strategy with other ones.

Research Background

The Impacts of Geographic Dispersion of Team Members on (OSS) Team Performance

In this study, we categorize the geographic dispersions of team members into two settings: geographically distributed and co-located ones. Geographically collocated team members often reside or work in the same geographical location. The level of location often depends on the lens of the observer and the application context. It could be a building, a street or a city. In the context of OSS development, we generally consider OSS team members who are within the same city as geographically co-located. By this definition, such co-located team members are in closer proximity with each other than geographically distributed OSS team members (i.e., developers who are not in the same city). Thus, geographically co-located OSS team members has much lower costs and in general more likely to have face-to-face meetings than distributed members.

In most OSS developer teams, both settings often co-exist, effectively creating three situations. When the majority team members are in the same location (city), this team can be in general considered as a co-located team (Sharp et al. 2012). An extreme example can be a group people work in the same level of a building. On the other hand, if the majority are distributed in multiple locations (cites), it can be considered as a distributed team. An extreme case can be a small group of developers without any two of them are in the same country. However, previous research on OSS (Crowston et al. 2005; Crowston and Scozzi 2002) and global software development (GSD) teams(Jalali and Wohlin 2012; Šmite et al. 2010) found that most such teams have a mixture of both settings of members and can be called hybrid teams (Staples and Webster 2008).

Moreover, both types of team members use face-to-face interactions and virtual communication technologies (e.g., online voice chat) to collaborate with each other. However, comparing with geographically distributed team members, the co-located team members have the capabilities and are more likely to engage in face-to-face interactions.

Previous virtual team studies often focused on the impacts of an individual setting of the geographic dispersion alone – distributed or co-located – on team performance (Hambley et al. 2007; O'Leary and Cummings 2007; Purvanova and Bono 2009; Staples and Zhao 2006), or compare the effects of virtual collaboration with face-to-face collaboration in these two settings separately. They overlooked the situation when the situation when both settings (and collaboration styles) coexist in the same hybrid team. Our unique empirical data set from the open source software (OSS) development community provides us an ideal environment to study the impacts of these two settings in such a mixture situation since most OSS project teams are hybrid teams(Crowston et al. 2005).

Geographically Co-located Setting

The majority of existing studies on face-to-face collaborations focused on teams in traditional organizations in which most members are co-located in a workplace. This is mainly because geographically co-located team members work in close physical proximity and are more likely to have frequent face-to-face interactions. But Crowston et al. (2005) suggested that face-to-face collaborations also existed in many OSS project teams which many studies consider as a largely distributed, virtual collaboration oriented environment. This means both settings may co-exist in real world OSS project teams and significantly affect their performance and success at the same time. However their relative impacts are rarely investigated in such situations.

It was found that face-to-face collaboration in general allow people observing each other’s physical and emotional cues which can improve team members’ innovation abilities and contribution motivations (Hart 2001). It was also found that the social and interpersonal cues that provide the basis for trusts are lacking in geographically distributed teams (Branson et al. 2008). Moreover, Face-to-face interactions can help provide a highly self-organizing work style, as well as a creative and innovative working environment where team members are highly committed (Branson et al. 2008). Staples (2006) found that geographically co-located team members share higher satisfaction of their teams.

Moreover, face-to-face collaborations can promote team productivity by providing more efficient and personal communications among team members, and thereby improve team performance. Olson et al. (2002) found that “when people have questions, often the person who could answer it (i.e., a fellow worker who had more experience or expertise on a topic) was at hand”. Warkentin et al. (1997) investigated 11 geographically co-located teams and discovered that they have high degree of cohesion. In addition, it was found that task conflicts among team members rarely existed in geographically co-located teams(Staples and Zhao 2006; Wakefield et al. 2008).

Geographically Distributed Setting

On the other hand, virtual collaboration also plays an important role in team work. Such collaboration often rely on CMC based technologies such as internet, email or online video conferencing. Virtual collaboration can bring team members effective connecting, better time management, and fast team reconfiguration (Lipnack and Stamps 1999; Townsend et al. 1998). Townsend et al. (1998) found that the main advantage of virtual collaboration is its flexibility to efficiently connect geographically distributed team members, which can largely improve team performance and effectiveness. Geographically distributed team members usually work in a 24-hour cycle, and can manage their working time in a more flexible and efficient style (Lipnack and Stamps 1999). Thus Bergiel et al. (2006) suggested that virtual collaboration can transcend the boundaries of time, distance, and organizations, thereby improving team members' creativity and efficiency.

However, despite of the above advantages virtual collaboration can bring to improve team performance, they cannot provide physical and emotional cues which promote team members' innovation and motivation (Hart 2001). Mihhailova et al. (2007) found that geographically distributed team members generally lack visual contact, frequent interactions and fast feedback. Such teams often face greater obstacles towards maintaining shared goals due to their large physical distance and therefore often tends to fall apart from each other. The lack of common work locations and cultures often disrupts a geographical distributed team's mutual awareness of individual members (Cramton 2001). It has also caused that team members are unable to find interpersonal cues which are crucial to mutual understanding and effective interactions (Dubé and Paré 2004). In addition, geographically distributed teams are more likely to experience greater conflicts than co-located teams (Wakefield et al. 2008) and thereby result in decreased team productivity and satisfaction(Hambrick et al. 1998; Lau and Murnighan 1998).

Spatial vs. Temporal Geographic Dispersion

Moreover, there are two aspects of geographic dispersion for team members: spatial and temporal. Previous studies mainly focused on the spatial geographic dispersion measured by the average geographic distance among team members. They have found a positive association between physical proximity and interpersonal communication efficiency (Festinger et al. 1950; Kiesler and Cummings 2002).

On the other hand, it was also found that temporal dispersion is negatively correlated with team performance (Griffith et al. 2003; Martins et al. 2004). A team's temporal geographic dispersion is often measured by the time zone differences among team members(O'Leary and Cummings 2007). Team members who are distributed in distant time zones, usually share fewer work hours and need to put more efforts in synchronize their working time with others considerably. Because of the lack of synchronous collaboration or the high costs for doing so, greater temporal dispersion often hinders real-time problem solving and spontaneous communications(Burke et al. 1999; Dennis and Kinney 1998) .The difficulties of synchronize communications grow as the temporal dispersion increases among team members (Davison et al. 2006).

Then a major research challenge is to distinguish the impacts between spatial geographic dispersion and temporal geographic dispersion on team performance (OSS project success). This is mainly because in real world many team members are dispersed both in the small spatial and temporal distances(O'Leary and Cummings 2007). In such situations, it is difficult to find out the good performance of a team is due to face-to-face collaboration enabled by close physical proximity enabled by being in the same time zone.

Research Model

To summarize, previous research on the impacts of the geographical dispersions of team members on team performance focused on either co-located or distributed setting alone, largely overlooked the situation where both settings co-exist, like the OSS development environment. Moreover, the impacts of spatial and temporal geographic dispersion is not clearly distinguished since most co-located team members are also in the same time zone.

Our study adopted Media synchronicity theory (MST) proposed by Dennis et al. (2008) to conceptually explain how geographic dispersion affect project success in our research model as Figure 1 shows. MST suggests that synchronicity exists among workmates at the same time with a shared focus. It defines media synchronicity as the "the extent to which the capabilities of a communication medium enable individuals to achieve synchronicity"(Davison et al. 2006; Dennis and Valacich 1999).

Geographic Dispersion

By definition, face-to-face communication is synchronistic. It occurs when the sender and receiver are in the same place at the same time. Comparing with other media forms such as virtual collaborations, this form is the richest media since it allows for both the sender and receiver to observe cues such as facial expressions, tone of voice, and other body language, all of which add context to the verbal message (Daft and Lengel 1986). Face-to-face communication allows multiple symbol sets to be transmitted simultaneously. Physical, visual, and verbal symbol sets are fast to encode, facilitating turn-taking and coordination and making interactions faster (Goffman 2005; Williams 1977). Therefore, face-to-face interactions as a media form that uniquely incorporates these symbol sets have greater capability to support synchronicity and result in better communication performance than virtual collaborations. Our research model then suggests that lower geographic dispersion (which induce more likelihood of face-to-face interactions) may improve team performance by enhancing collaboration efficiency and build trustful relationships.

Moreover, the successful completion of complex tasks often involves synchronous collaborations among team members. Therefore collaboration performance may be improved when individuals in the same time zone can better coordinate and communicate because of the synchronicity (Dennis et al. 2008). But we hypothesize that the current computer-mediation based virtual collaboration technology can't fully replace the role of face-to-face collaborations. Teams which use both the face-to-face and virtual collaboration appropriately may many have better performance.

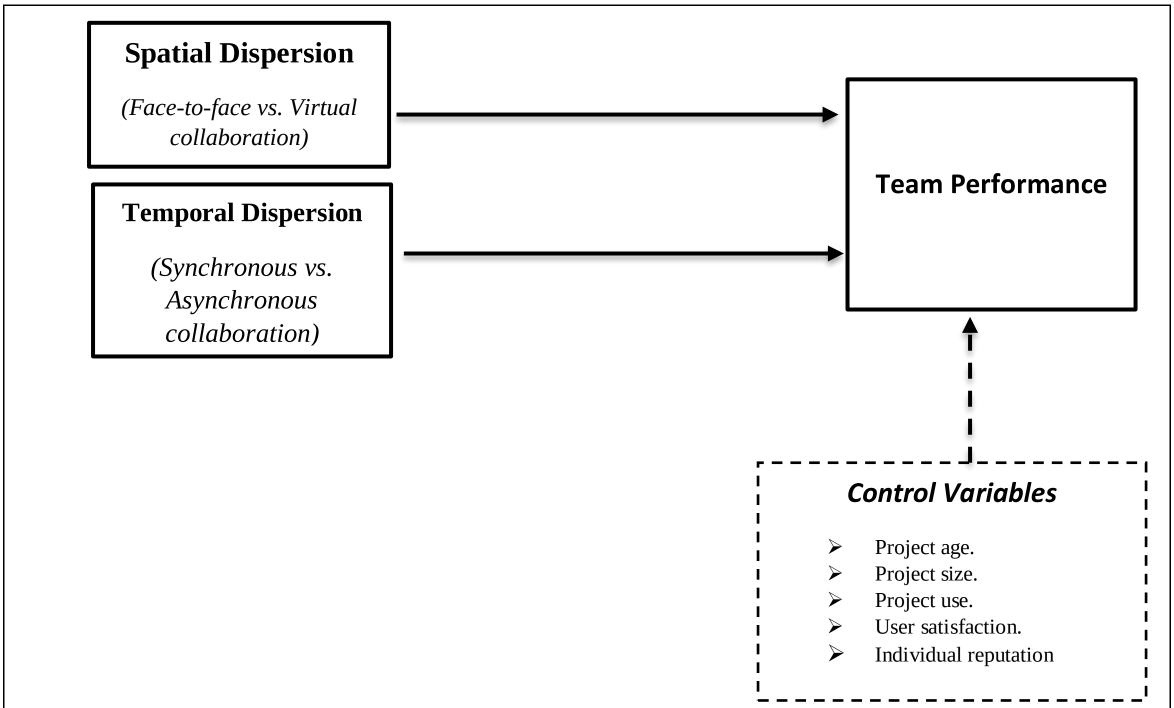

Figure 1. Theoretical Framework

To summarize, our research model in Figure 1 generally suggests that lower spatial dispersion (likely more face-to-face interactions) and lower temporal dispersion can lead to better team performance mainly due to two MST based mechanisms. First, lower spatial dispersion with more face-to-face interactions provides the team members the richest media to communicate and collaborate, thereby improving their collaboration efficiency and build trustful relationships and environment in the long run. Second, the synchronicity brought by being in the same time zone and various virtual collaboration technologies can improve collaboration efficiency among team members. But the effects of these two mechanisms in this model need to be distinguished, along with various project related control variables we extracted from relevant studies as shown in Figure 1.

Research Questions and Hypotheses

As discussed in the previous sections, previous studies mainly studied the associations between geographic dispersions and team performance in individual settings of geographic dispersion. In our study, we adopt instrumental variable estimation methods to investigate the causal effects of geographic dispersion on team performance (project success) in a large scale, real world OSS development community where both geographic dispersion settings co-exist in many projects. So our first research question and hypothesis are as follows:

RQ1: What are the impacts of geographic dispersion of OSS team members on their team performance (project success)?

H1: The average geographic distance among the members of an OSS project team, negatively affects the rating of this project.

Moreover, existing research has not distinguished the impacts between spatial (face-to-face collaboration) and temporal (convenience factors) geographic dispersion on OSS team performance. Therefore, the second research question is:

RQ2: Do spatial geographic dispersion and temporal dispersion have different effects on OSS team performance (project success)?

H2: The rating of an OSS project is negatively affected by the average geographic distance among its team members after controlling for the average time zone difference of these members.

Research Design

We developed a set of Java programs to automatically query and retrieve data through the API provided by Open Hub. Since all retrieved data items are in XML format, a parser program was developed to parse them into a database. Such information includes OSS developers' project participation and contributions, their positive evaluation choices, location, nationality, programming language preferences, development activities, and project statistics. Moreover, it also keeps track of the changes in the source code of each listed OSS project from the version control systems and calculated software metrics such as the total number of changes (i.e. commits) at different time periods.

In addition, Open Hub also encourages OSS developers to register, edit and discuss their projects in its online social-networking community. Thus it provides two unique types of social network information which is not available at other major OSS portals including Sourceforge.net. The first type is from Open Hub's evaluation mechanism – the "Kudo" links. Each Open Hub community member can send another member a link called a "Kudo" which shows a gesture of appreciation, praise, or endorsement from the sender to the receiver. An Open Hub developer may appreciate another developer's OSS-related characteristics (e.g., source code, code comments) and sends a Kudo link to him or her as recognition. He may also appreciate the contributions of his or her co-developer in an OSS project and sends a Kudo link to this co-developer. Moreover, a developer may receive one-to-one help from another developer in OSS development work and thus sends a Kudo link to express his or her thanks. Therefore, a Kudo link represents a positive evaluation from the sender to the receiver.

Most importantly, Open Hub allows developers to register their addresses in their database and parse them to coordinates (with longitude and latitude information). Moreover, Crowston et al.

(2005) found that many OSS teams are most hybrid teams in which co-located and distributed team members co-exist. Therefore, such location information along with the project rating provides us a unique opportunity to investigate the relationship between the average team members' distance (representing different geographic dispersion patterns) and team performance (project rating).

Measurement

Dependent Variable: OSS Project Success (Team Performance)

In our study, we adopt the average rating score of the users of an OSS project in Openhub platform as the measurement for the success of that project. Any user registered in Openhub can rate an OSS project based on her overall satisfaction with a rating score ranging from 1.0 to 5.0 (e.g., 3.6), in which 5.0 is the best rating. We use the average rating score of an Openhub OSS project as our main measure of project success (team performance). In our empirical analysis, the average rating score is used as the dependent variable and denoted as **RAT**. Table 1 shows the detailed definitions of variables used in our empirical analysis.

Independent Variables: Measuring the Level of Spatial and Temporal Geographic Dispersion (for an OSS Project)

In addition, Openhub provides the location information of project team members, as often called project contributors. The location information includes three perspectives: 1) an optional text description of this contributor's claimed location; 2) the country code – a string (e.g., US) representing his country; and 3) a pair of floating point values representing the latitude and longitude of his exact (home or work) location. The location data has been validated by either the Google geocoder web service.

Most studies measure the spatial dispersion by a dichotomous variable to distinguish team processes and outcomes between collocated and distributed teams (Ehrlich et al. 2008;

Herbsleb and Mockus 2003). Some studies also consider the number of sites as the degree of spatial dispersion (Nguyen et al. 2008; Ramasubbu et al. 2011). While others examined whether team members were located in the same room or different building, city, or country (Chudoba et al. 2005; Espinosa et al. 2012).

In our study, we use the latitude and longitude values of members in an OSS project for calculating the average geographic distance (denoted as **GeoDist**) and average time zone difference (denoted as **TimeZoneDist**) among these members, in order to measure the level of spatial and temporal dispersion of this project. They are used as independent variables in our empirical analysis.

Control Variables

The control variables are project level factors. Many of them have been studied in previous research on OSS project success. In our analysis, these variables are as control variables. They include the total lines of codes, total time and man power efforts invested to a project by all team members, total number of participating contributors (team members), etc.

Table 1. Variable Definitions (All the variables are measured in the project level)	
Dependent Variable	
RAT (Average Rating Score)	It is a score ranging from 1.0 to 5.0, representing the average value of all ratings by Openhub community members. 1.0 is the lowest possible rating and 5.0 is the highest possible rating.
Independent Variables	
GeoDist (Average geographic distance)	The average value of the geographic distance of all the contributor pairs of the project. This value is measured by kilometers. The calculation formula is as follows: $GeoDist = \frac{\sum_{i,j=1}^{N} dist(d_i, d_j)}{(N^2 - N)/2}$ • In the formula, $dist(d_i, d_j)$ is the great circle distance between the two contributors d_i and d_j . • N is the total number of contributors of the project.

TimeZoneDist (Average time zone difference)	The average value of the time zone difference of all the contributor pairs of the project. The calculation formula is as follows: $$TimeZoneDist = \frac{\sum_{i,j=1}^{N} timediff(d_i, d_j)}{(N^2 - N)/2}$$ • In the formula, *timediff* is the time zone difference between the two contributors d_i and d_j . For example, if the contributor d_i is located in time zone T1 (UTC+00:00 i.e., London), while the contributor d_j is located in time zone T2 (UTC-05:00 i.e., New York). • The $timediff(d_i, d_j)$ is the absolute value of the time zone difference between the two contributors A and B: $timediff(d_i, d_j)$. If the value of *timediff* is larger than 12, we subtract it by 24 to keep it between 0 and 12. • N is the total number of the contributor pairs of the project.
Control Variables	
AGE (Project age)	Openhub platform provides the created and latest updated time for each project. The number of calendar weeks between the created and latest updated time stamp of the project.
NUM_CON (Total number of contributors)	The total number of contributors of the project.
NUM_USR (Total number of users)	The total number of users of the project.
AVG_COMMITS	Average number of commits per contributor in the project.
AVG_MMONTH	Average number of efforts (months) invested by all contributors of the project.
AVG_LOC	The average number of lines of code of the project, excluding comments and blanks.
AVG_KUDO	The average value of the Kudo rank per contributor in the project. Kudo rank is a score which represents the reputation level of the contributor in the Openhub platform.
RAT_PER	The proportion of the users of the project who rated it.
REW_PER	The proportion of the users of the project who reviewed it.

Empirical Model and Instrumental Variable (IV) Estimation

In our study, we adopt a linear mode with instrumental variable estimation to investigate the causal relationship between geographic dispersion and project success. We use the two-stage least squares (2SLS) procedure for the instrumental variable estimation.

Instrumental Variable (IV) Estimation

In order to investigate the causal effects of geographic dispersion on project success, we develop an instrumental variables to examine the possible endogeneity our main independent variable - average geographic distance – may bring.

The contributors/team members of an OSS project P may also participate other projects. We use the average value of the average geographic distances for all these other projects (for all team members) can be used as the instrumental variable. This is mainly because that a contributor's main preferences to work with either co-located or distributed team members is usually consistent across the project he participated. So the independent variable (i.e., average geographic distance of P) should be correlated with the instrumental variable (i.e., the average value of average geographic distances for all other projects P's members have participated).

On the other hand, in general those other projects' successes (average rating scores) should not be correlated with P's success because they do not systematically share anything else except one common developer. This is especially true for large projects with many team members. Then the selected instrumental variable is correlated with the independent variable, but unrelated with the dependent variable.

The instrumental variable $ivdist_p$ can be calculated using the following equations:

$$ivdist_p = \frac{\sum_{i=1}^{N} Dist_i}{N} \quad (1), \quad Dist_{i=} \frac{\sum_{j=1}^{M} GeoDist_j * comrate_j}{M} \quad (2), \quad comrate_j = \frac{commits_j}{total\_num\_commits} \quad (3)$$

where N is the total number of contributors of the project p, and $Dist_i$ is the average value of average geographic distance for the outside projects team member i has participated before. M is the total number of projects that i has participated before except the project p. $GeoDist_j$ is the average geographic distance of one of i's outside projects j, and $comrate_j$ is the share of i's contributions on j ($commits_j$) comparing with his total contributions (total_num_commits).

Summary Statistics

Table 2 reports the summary statistics of our empirical data. The mean value of **GeoDist** is 4241, and the maximum is 19857.4. The mean value of **TimeZoneDist** is 3, and the maximum is 12. The mean value of **RAT** is 3.8 and the maximum is 5. In our regression analysis, we use logarithmic function to transform the variables. Both the GeoDist and TimeZoneDist are negatively correlated with the average rating score of the project (RAT).

It was found that the **GeoDist** is highly positively correlated with the **TimeZoneDist**. It is understandable since most co-located team members should be in the same or adjacent time zones. It confirms the necessity to distinguish the impacts between the spatial and temporal geographic dispersion on OSS team performance (project success).

Table 2. Summary Statistics (Variables Definitions are in Table 1)

	RAT	*Geo-Dist*	*Time-ZoneDist*	*AGE*	*NUM-CON*	*NUM-USR*	*AVG-COMMITS*	*AVG-MMONTHS*	*AVG-LOC*	*AVG-KUDO*	*RAT-PER*	*REW-PER*
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
Mean	4.800	4241	3	349	113.1	6	68.68	10.63	13492	0.533	0.014	7.932
SD	0.546	3684.729	2.678	100.546	569.833	436.519	213.636	18.684	79241.410	0.366	0.079	0.925
Bivairate Correlations												
(1)	1											
(2)	-0.087	1										
(3)	-0.083	0.794	1									
(4)	-0.288	0.092	0.091	1								
(5)	-0.103	0.185	0.178	0.088	1							
(6)	-0.276	0.159	0.160	0.380	0.346	1						
(7)	-0.015	-0.075	-0.071	0.159	-0.056	0.130	1					
(8)	-0.037	-0.119	-0.117	0.214	-0.210	0.078	0.594	1				
(9)	-0.002	-0.099	-0.097	0.136	-0.167	0.067	0.489	0.553	1			
(10)	-0.003	-0.033	-0.027	-0.011	-0.188	-0.041	-0.056	0.011	-0.038	1		
(11)	0.116	-0.126	-0.134	-0.145	-0.232	-0.487	-0.017	0.011	0.016	-0.022	1	
(12)	-0.120	0.056	0.050	0.188	0.106	0.235	0.104	0.085	0.079	-0.045	-0.074	1

Results and Findings

Table 3 shows the regression results of our main empirical analysis using the instrumental variable estimation method. We examine the causal impacts of the average geographic distance among OSS team members on the project average rating score. The coefficient on GeoDist estimates the magnitude of its average effects on the project success. The negative sign of GeoDist indicates that the smaller the average geographic distance a project has, the higher the average rating score it is likely to get. This indicates that, in general, higher likelihood of face-to-face collaboration among the members of an OSS project will increase the chances of its success. After controlling for the **TimeZoneDist**, the **GeoDist** is still significant with a negative sign. Therefore both hypotheses are supported by the above results.

Table 3 also shows that the **AGE** is a significant variable with negative sign. It suggests that, if everything else is the same, the younger a project is, the more successful may become. It is consistent with previous OSS research that found the younger projects are more likely to attract developers to join in. The coefficient of the **NUM_CON** is positive and significant. It indicates that the larger number of contributors a project has, the higher the average rating score of the project is likely to have. The **NUM_USR** is significant with negative sign, suggesting that the smaller number of users a project has, the higher the average rating score of the project is likely to have. The coefficient of **AVG_COMMITS** is positive and significant. It means that the more commits a project gets, the higher the average rating score of the project is likely to have. The **AVG_MMONTH** is a significant variable with negative sign, suggesting that the smaller the man power is invested into a project, the higher the average rating score the project is likely to have.

Table 3. Regression Results of Main Analysis		
Independent Variables		
	Without TimeZoneDist	***With TimeZoneDist***
GeoDist	**-0.0598 *** (0.0127)**	**-0.1204 *** (0.0279)**
TimeZoneDist	NA	0.2473 *** (0.0593)
Control Variables		
AGE	-0.1493 *** (0.0199)	-0.1500 *** (0.0204)
NUM_CON	0.0253 ** (0.0084)	0.0242 ** (0.0085)
NUM_USR	-0.0552 *** (0.0074)	-0.0653 ** (0.0075)
AVG_COMMITS	0.0553 *** (0.0107)	0.0539 *** (0.0109)
AVG_MMONTH	-0.0632 *** (0.0161)	-0.0497 ** (0.0160)
AVG_LOC	0.0070 (0.0055)	0.0083 (0.0056)
AVG_KUDO	-0.0198 (0.0614)	0.0194 (0.0641)
RAT_PER	-0.1421 *** (0.0426)	-0.1356 ** (0.0430)
REW_PER	-0.1691 (0.1189)	-0.1138 (0.1231)

Note: The standard errors are reported in parentheses.
P values are as follows: *p < 0.10; **p < 0.05; ***p < 0.01.

Regression Results for Different Time Zone Intervals

We divided all projects into two groups according to the TimeZoneDist Values: **TimeZoneDist**<=1 vs. **TimeZoneDist** >1. We would like to examine the difference in regression results for the projects whose members are within the same (and adjacent) time zone or not. Table 4 reports the regression results in the two groups respectively. We found that GeoDist is still significant for projects that TimeZoneDist<=1 with a smaller negative coefficient. This indicates the positive effects of close proximity (larger likelihood of face-to-face collaboration) on team performance is stronger for members within the same time zone. However, for projects whose average distance is large (since their members are not in the same adjacent time zones). The effects of spatial distance is not significant on project success. We may need to further analyze the impacts across different distance intervals in our future study.

Table 4. Regression Results for Different TimeZoneDist Intervals		
	TimeZoneDist<=1	***TimeZoneDist>1***
Independent Variables		
GeoDist	**-0.0855 *** (0.0281)**	-12.4135 (1.5106)
TimeZoneDist	0.3956 *** (0.1327)	14.1712 (1.5287)
Control Variables		
AGE	-0.1101 *** (0.0310)	0.1280 (0.0512)
NUM_CON	0.0557 ** (0.0177)	0.0139 (0.0117)
NUM_USR	-0.0587 ** (0.0201)	-0.0652 (0.0131)
AVG_COMMITS	0.0542 ** (0.0200)	0.0311 (0.0184)
AVG_MMONTH	-0.0519 (0.0288)	-0.1884 (0.0344)
AVG_LOC	0.0109 (0.0097)	0.0220 (0.0103)
AVG_KUDO	0.0987 (0.1051)	-0.7757 (0.1473)
RAT_PER	-0.0627 (0.0768)	0.3232 (0.0930)
REW_PER	0.0811 (0.1924)	-0.6413 (0.2010)

Regression Results for Different Project Sizes (Number of Contributors/Team Members) Intervals

We then divided the projects into two groups based on the total number of contributors (team members) of a project has: **NUM_CON**<=20 and **NUM_CON**>20. In our data sample almost 45% projects has less than 20 contributors. We also divided projects into two groups based on the total number of contributors (project size) of a project has:**NUM_CON**<=70 and **NUM_CON**>70.

Table 5 reports the regression results for these two groups respectively. The results show that for smaller projects (**NUM_CON**<=20 or **NUM_CON**<=70), both hypotheses hold. However, for very large projects (**NUM_CON**>70), spatial geographic dispersion has no significant impacts on team performance.

Table 5. Regression Results for Different Project Size

	NUM_CON<=20	*NUM_CON>20*	*NUM_CON<=70*	*NUM_CON>70*
Independent Variables				
GeoDist	**-0.1327 *** (0.0321)**	**-0.1561 ** (0.0488)**	**-0.1310 *** (0.0288)**	-0.0281 (0.0811)
TimeZoneDist	0.2310 * (0.0744)	0.2599 ** (0.0892)	0.2854 *** (0.0640)	0.0041 (0.9730)
Control Variables				
AGE	-0.1388 *** (0.0283)	-0.1742 *** (0.0265)	-0.1563 *** (0.0243)	-0.1508 *** (0.0336)
NUM_USR	-0.0660 *** (0.0139)	-0.0516 *** (0.0078)	-0.0768 *** (0.0101)	-0.0550 *** (0.0103)
AVG_COMMITS	0.0508 ** (0.0606)	0.0725 *** (0.0572)	0.0519 *** (0.0504)	0.0683 *** (0.01960)
AVG_MMONTH	-0.0630 * (0.1543)	-0.0731 *** (0.1637)	-0.0571 ** (0.1316)	-0.0371 (0.0259)
AVG_LOC	0.0102 (0.0240)	-0.0009 (0.0193)	0.0115 (0.0196)	0.0009 (0.0105)
AVG_KUDO	0.1150 (0.0153)	-0.1462 (0.0141)	0.0739 (0.0128)	-0.1800 (0.1315)
RAT_PER	-0.1115 (0.0079)	-0.1714 ** (0.0075)	-0.1383 ** (0.0067)	-0.2368 ** (0.0884)
REW_PER	0.0380 (0.0848)	-0.4732 ** (0.0891)	-0.0883 (0.0724)	-0.4929 (0.4388)

Regression Results for Core Project Contributors

Previous OSS research there may be a group of core developers that have significant impacts on the success of this project. Therefore, we also analyzed the impacts of average geographic distance of the core developers in an OSS project team on its success (average rating score). We define and select the core contributors of an OSS project using the following procedure:

1): Calculate the share of the developer's commits for a project as: $comrate = \frac{commits}{project\_total\_commits}$

where $commits$ represents the number of the commits this developer contributes to this project and: $project\_total\_commits$ refers to the total number of commits this project ever received.

2): For each project, we then select the top 20% of developers ranked by their $comrate$ as the core developers. The rest are regarded as periphery developers of the project. We calculate the average geographic distance and average time zone difference of these core developers and analyze their impacts on the project success (rating).

Table 6 shows the regression results for three different groups: core, periphery and core-periphery. The signs of the GeoDist are all negative across all three groups. But the GeoDist is only significant in the core group. It indicates that the average geographic distance only matters in the core group. Meanwhile, the coefficient of the TimeZoneDist is also only significant in core group, suggesting that the TimeZoneDist only matters in the core group.

Table 6. Regression Results of Different Project Contributor Groups

	CORE	PERIPHERY	CORE-PERIPHERY
Independent Variables			
GeoDist	**-0.1223 *** (0.0225)**	-0.0479 (0.0311)	-0.0742 (0.0405)
TimeZoneDist	0.2270 *** (0.0463)	0.0890 (0.0572)	0.1292 (0.0780)
Control Variables			
AGE	-0.1692 *** (0.0208)	-0.1719 *** (0.0362)	-0.2078 *** (0.0254)
NUM_CON	0.0330 *** (0.0095)	0.0313 (0.0250)	0.0016 (0.0091)
NUM_USR	-0.0448 *** (0.0083)	-0.0531 *** (0.0094)	-0.0578 *** (0.0080)
AVG_COMMITS	0.0383 ** (0.0122)	0.0767 *** (0.0151)	0.0732 *** (0.0132)
AVG_MMONTH	-0.0440 * (0.0172)	-0.0425 (0.0220)	-0.0605 ** (0.0190)
AVG_LOC	0.0055 (0.0060)	-0.0047 (0.0083)	0.0018 (0.0896)

AVG_KUDO	0.0242 (0.0710)	-0.1290 (0.1134)	-0.0875 (0.0520)
RAT_PER	-0.1019 * (0.0454)	-0.1516 * (0.0610)	-0.2051 *** (0.1759)
REW_PER	-0.1185 (0.1299)	-0.4939 * (0.2077)	-0.2801 (0.1113)

Findings

Our research has implications for OSS contexts in several perspectives. First although numerous studies have examined the OSS project's success, the impacts of geographic dispersion on OSS performance has not studied with rigorous econometric methods. Furthermore, we distinguish the difference of the impacts between spatial geographic dispersion and temporal geographic dispersion. Our research findings show that the spatial geographic dispersion has a negative impact on project success, even after controlling temporal geographic dispersion.

Second, geographic dispersion can be regarded as a major actionable factor for project leaders, especially in commercial software development projects. Although previous studies pointed out the role of face-to-face collaborations in virtual teams especially in OSS teams. There is no rigorous empirical support for this practice, furthermore the causal effects of geographic dispersion on team performance are rarely explored.

Third, this study has useful research implications for the virtual team context. Virtual team literature has views of the impacts of geographic dispersion on the team performance (Cramton 2001; O'Leary and Cummings 2007), but without proper empirical evidence especially for the causal relationship between them (O'Leary and Cummings 2007). This study provides solid, empirical evidence about the relationship between the geographic dispersion and virtual team performance in the OSS contexts.

Practical Implications

Based on the above findings, we offer a set of suggestions for both individual OSS developers and project managers about managing project members, building trusting environment and facilitate good influence, and thereby improving team performance (project success). For individual developers, our suggestions focus on increasing the visibility of their locations on Open Hub community, as well as their previous success, programming skills, and positive evaluations (kudos), aiming to increase their chances to be discovered by more local developers and subsequently build trust and facilitate efficient collaborations.

For project managers, our suggestions aim to help them to build an environment that can facilitate more face-to-face interactions and induce trustful relationships among co-located team members. Such trustful relationships then can facilitate more efficient collaborations, and thereby improve team performance and project success. Moreover, such an OSS project with such an environment may attract more local project developers through their desires to achieve personal success by project participations.

More specifically, the project managers can provide more face-to-face interaction opportunities for local developers within and outside the project group, aiming to induce stronger personal relationships and build a more trustful and collaborative environment. Such opportunities may include but not limited to: 1) organizing social events for the team members in the city where most of them live in, especially for the core members, 2) organizing development workshops for co-located members to improve programing skills or develop future develop plans, and 3) assigning complex group tasks to co-located developers or developers within the same time zone. In particular, we would like to take the advantage of synchronous collaborations for developers within the same time zone. Co-located developers is good at facilitating trustful relationships and improve communication and collaboration efficiency in general. Developers within the same of adjacent time zones are good at coordinating each other to solve real time complex problems. These actions all together may induce more face-to-face interactions, synchronous problem solving, and facilitate various collaborative and trustful relationships among co-located developers.

Conclusion, Limitations and Future Research

Our results show that the average geographic distance (representing the likelihood of face-to-face interactions) negatively affects the team performance/project success, even controlling for the temporal geographic dispersion. We have also distinguished the impacts of the spatial and temporal geographic dispersion on team performance. Since project members in the same time zone may also be in cities that are further apart (e.g., Washington and Miami). To further exclude the effects of being in the same time zone, we use IV estimation to further analyze the relationship between average geographic distance and the rating of projects in whose average time zone distance TimeZoneDist<=1. The results still show that OSS teams negatively affects the project success. Moreover, both analyses show that time zone dispersion also negatively affects project success. Then both of our hypotheses are supported.

As mentioned in the research background section, previous studies focus on the impacts of geographic dispersion of team members on team performance in either co-located or distributed setting alone(Crowston et al. 2005; Staples and Zhao 2006; Warkentin et al. 1997), while the situation in which both settings co-exists is rarely studied. OSS development community provides an ideal environment in which most teams are hybrid consisting both settings of geographic dispersions for their members. We have adopted rigorous econometric methods to empirically examine the causal effects of geographic dispersion on the OSS project success, especially in the situation where both geographically co-located and distributed team members co-exist in the same teams.

The contributions of our study are three folds: first, this research contributes to OSS and collaboration literature by enriching the understanding of the theoretical and empirical relationship between the geographic dispersions of team members and (OSS) team performance (project success) when co-located and distributed dispersion co-exist in same teams. Second, we distinguish the impacts of the spatial and temporal geographic dispersion on (OSS) project success. Meanwhile, this study provides empirical quantitative evidence to the findings of the qualitative study done by Crowston (2005). Our study support that face-to-face (f2f) collaboration in the OSS context can achieve better project success from a quantitative perspective. Finally, our empirical findings may provide useful insights and for OSS stakeholders to devise practical strategies to improve their team performance and project success.

We then discuss the limitations of our study and the directions for future research. First, our empirical results about the relationship between geographic dispersion and project success is consistent with our MST based hypotheses. While this is reasonable, the suggestions we made to both individual developers and project managers for increasing face-to-face interaction opportunities and its impacts on project success is only conceptually analyzed and have not been empirically tested. Future research using experiment or survey methodology should be conducted to validate these suggestions and further improve our understanding between geographic dispersion and OSS project success.

Another limitation is that several developer characteristics such as gender and race are not are not available in Openhub dataset and thus not included in our study. Thus an important extension of this paper would be studying how different developers' geographic dispersion affect project success when such personal data becomes available.

To summarize, our future work mainly consists of three research directions, including 1) empirically validating the effectiveness of our suggestions for inducing face-to-face interactions and building trustful relationships among local developers, 2) investigating how different developers' geographic dispersion affect project success, and 3) studying the peer effects among co-located developers of the project and their impacts on project participations and contribution.

This page is intentionally left blank

Part 3: Study II: The Impacts of Online Social Influence on OSS Project Participation

Understanding the Impacts of Social Influence on Initial and Sustained Participation in Open Source Software Projects

Xiao Li
University of Zurich
Zurich, Switzerland
davidlixiao@ifi.uzh.ch

Xuan Yang
Hong Kong Shue Yan University
Hong Kong S.A.R., China
sherry.teen@gmail.com

Daning Hu
University of Zurich
Zurich, Switzerland
hdaning@ifi.uzh.ch

Ji Wu
Business School, Sun Yat-sen University
Guangzhou, China
wugide@gmail.com

Harry Jiannan Wang
University of Delaware
Newark, USA
The Chinese University of Hong Kong, Shenzhen, China
hjwang@udel.edu

This study is based on the submission to: Proceedings of the 38th International Conference on Information Systems (ICIS 2017), Seoul, South Korea

Abstract

Nowadays open source software (OSS) development platform are increasingly using social networking-like functions such as microblogging, aiming to use developers' social influence to attract more high quality project participation. However, social influence is largely overlooked in OSS participation research and has often been studied from an economic utility framework in existing literature. Such a framework may be not suitable for analyzing the often non-monetary motivations behind OSS developer participations. We plan to empirically investigate the impacts of word of mouth (WOM) and observational learning (OL) on OSS developers' initial and sustained participation behaviors, using data from a large OSS platform with more than 669,000 projects. Our preliminary results show that social influence has significant but rather different impacts on initial and sustained OSS participation. Specifically, the impacts of WOM on developers' sustained participation faded away after initial participation as they can better evaluate the underlying project and its members' opinion.

Keywords: Open source software development, Participation, Social influences, Social Coding

Introduction

Nowadays major open source software (OSS) development platforms like GitHub and Open Hub are providing social networking - like functions such as microblogging feeds, followers, and reputation systems to allow individual developers to track the participation activities of their peer developers and the progress of specific projects (Dabbish et al. 2012). Through such social media technologies, OSS developers can publish their opinions and activities of specific projects online to influence others' participation decisions. Since OSS project success largely relies on developers' voluntary and sustained participations (Fang and Neufeld 2009; Roberts et al. 2006), it is crucial for stakeholders to better understand the impacts of social influence on project participation behaviors, and develop strategies or functions accordingly to facilitate more participations and achieve better project success.

Social influence has been conceptualized as "how the behaviors of one's peers change the utility one expects to receive from engaging in a certain behavior and thus the likelihood that one will engage in that behavior" by (Aral 2011). It has been extensively studied in the economics and marketing literature (Arndt 1967; Bandura 1971; Katz and Lazarsfeld 1966). But these studies often employ an economic utility framework with cost-benefit analysis to examine the impacts of social influence on one's purchase decisions. On the other hand, OSS developers' participations are usually motivated beyond pure economic considerations, such as altruism, learning, and desire to gain community reputation(Hertel et al. 2003; Raymond 2001; Stewart 2005) (Hertel et al. 2003; Raymond 2000; Stewart 2005). Therefore, social influence may have quite different impacts on OSS project participations than the previously well-studied consumer adoption behaviors in the marketing and economic literature. However, to the best of our knowledge, little research has explored the impacts of social influence on OSS participations.

Moreover, prior OSS studies found that project success largely relies on two types of developers' participation behaviors - initial participation and sustained participation (i.e., continuous contribution) (Fang and Neufeld 2009; Roberts et al. 2006). We conjecture that, as a developer knows more about a project and its members when he is progressing from the initial participation stage, she can better evaluate the social influence (information) received from existing project members, and make more informed decisions on whether to contribute more to this project (i.e., sustained participation). Thus social influence may have differential impacts on developers' initial and sustained OSS project participation. However, such differences were rarely studied, especially from the perspective of social influence.

In this study, we intend to investigate the impacts of social influence on both types of OSS developers' participation outside the economic utility framework. We aim to gain insights on how to leverage different social influence mechanisms for improving participation and thereby achieving better project success. We collected real world data from a large OSS development platform called Open Hub (previous known as Ohloh.net). This platform provides detailed longitudinal data for about 669,000 projects and 3.85 million developers from 1987 to 2017. Its project detail pages also hosted a microblogging service and provide statistics of other project members' participation activities. This unique setting allows us to examine how WOM (microblogging feeds) and OL (others' participations) may affect developers' project participation decisions. Furthermore, the detailed information about developers' development activities in our data set enables us to study the differential impacts of social influence on developers' initial and sustained participation.

Our contribution is threefold. First, our study is among the first to adopt a social influence perspective to study OSS developers' participations. It can enrich the theoretical understanding of the impacts of social influence on individual behaviors that are not mainly geared towards to maximize economic utility. Second, our study investigated the possible differential impacts of the WOM and OL on OSS participations. Third, we can also examine how social influence may affect initial participation and the understudied "sustained" OSS participation. These findings may help to design better information systems or effective policies to improve both types of OSS participations.

Research Background

Social Influence and OSS Project Participation

Social influence have been well studied in economics and marketing (Arndt 1967; Bandura 1971; Katz and Lazarsfeld 1966). It is mainly used to explain socially connected individuals' correlated economic behaviors like product adoption. Thus these studies often adopt an economic utility framework with cost-benefit analysis to examine the impacts of social influence. For instance, previous marketing research suggested that consumers may depend on WOM or OL information to reduce uncertainty in their purchase decisions and thus avoid (obtain) potential economic loss (gain). However, such a framework is not suitable for studying OSS developers' participation behaviours. This is mainly because that prior OSS studies found that project participation is often driven by non-monetary motivations, such as software use(Franke and Von Hippel 2003; Hertel et al. 2003; Hippel and Krogh 2003; Von Krogh et al. 2003), altruism and fun(Hars and Ou 2001; Hemetsberger 2002; Lakhani and Wolf 2005; Luthiger and Jungwirth 2007), recognition and reputation (Hertel et al. 2003; Lakhani and Wolf 2005; Lerner and Tirole 2002; Raymond 1999; Roberts et al. 2006; Shah 2006; Von Krogh et al. 2003).

Word of Mouth and Observational Learning (on OSS Project Participation)

Two major types of social influence mechanisms were extensively studied in marketing and economic literature. First, Arndt (1967) defined the mechanism that consumers' product adoption is influenced by others' opinions and experiences as word of mouth (WOM). Nowadays, such opinions are often in the form of online reviews or social media communications. Second, Bandura (1977) and Barbagallo et al. (2008) defined the mechanism that individuals may observe and be influenced by others' actions without knowing the motives/reasons behind such actions as observational learning (OL). In this study, we also adopt this framework to study how these two mechanisms affect OSS developers' project participations.

Word of Mouth studies mainly focused on its impacts on consumer behaviors and product sales (Awad and Ragowsky 2008; Cheung and Thadani 2012; Rui et al. 2013).These studies indicated that WOM valence (positive or negative) can change consumers' evaluation of the products (Chevalier and Mayzlin 2006; Mizerski 1982), while WOM volume help facilitate better consumer awareness and increase the number of informed consumers. Moreover, the emergence of the Internet services like online reviews for "publicizing feedback and recommendations on products" has attracted many researchers to study WOM in the digital age (Chen and Xie 2008; Dellarocas 2003; Duan et al. 2008). For instance, Clemens et al. (2006) conducted a survey of online reviews from craft beer industry and found that products with high valence are likely to be bought again. Cheung et al. (2014) found that an increase in the volume of online product ratings can improve sales.

However, the impacts of WOM on OSS participations have not been well studied. Krishnamurthy (2003) suggested that in general there is a lack of resource for marketing OSS projects through traditional media. Then Bagozzi and Dholakia (2006) and Barbagallo et al. (2008) briefly discussed that WOM can be useful in advertising OSS projects and building awareness among developers. More recently, Santos et al. (2013) pointed out that WOM has great potential in influencing developers' participation behaviors. However, they all did not empirically investigate the impacts of WOM.

In the context of our study, microblogging service in the Open Hub community enables developers to disseminate their opinions, recommendations, and activities of OSS projects among developers. Jansen et al. (2009) and Hennig-Thusrau et al. (2015) suggested that microblogging offers a novel electronic channel of WOM. While earlier microblogging research mainly focuses on individuals' motivations to post (Davidson and Vaast 2009; Java et al. 2007; Zhao and Rosson 2009). More recent studies (Dabbish et al. 2012; Seebach et al. 2011; Tsay et al. 2012) have indicated that microblogging can enhance transparency and collaboration for software developers. We conjecture that using microblogging to publish developers' positive experiences, opinions, or participation activities can raise project awareness and in turn attract more participation.

Observational Learning research studies its impacts on consumer product adoption behaviors Bikhchandani et al. (1992). Their theoretical explanation suggests that OL information contains signals expressed by others' adoptions but not the reasons behind such actions. When there is limited product information, the publicly observed other consumers' adoptions by an individual outweighs her own private information in her adoption decision. As more and more consumers follow their predecessors' adoptions, an information cascade and behavior "herding" occur among people (Banerjee 1992).

The impacts of OL are amplified in online environment, as individual's online activities are becoming increasingly transparent (Cheung et al. 2015; Dellarocas et al. 2010; Ye et al. 2013; Zhou et al. 2013). For instance, Burke et al. (2009) found that social networking site users who see their friends' contributions are motivated to share more content. Dellarocas (2010) and Cheung et al. (2015) found OL affects people's information contributions in online communities.

Comparing with OL, the information conveyed through WOM are more of subjective (personal) opinions or evaluations. The publicly observed others' actions information is often in the form of objective statistics (e.g., sales). OSS developers' participation is a process which developers become more engaged and thus more familiar with the project and its members. We conjectured that developers' reliance on social influence may change in this process as they can better evaluate the project and its members' words due to such familiarity. Moreover, the difference in the objectivity of the information conveyed by OL and WOM may cause different changes in developers' reliance on social influence. The details of those changes will be examined in our study. In order to do that, we first review the literature about initial and sustained OSS participation, their differences, and how social influence may affect them differently.

Initial and Sustained OSS Project Participation

Existing OSS research on developers' participation behaviors mainly focused on the motivations of developers' initial participations (i.e., initial reasons for joining the projects) (Ghosh 2005; Hann et al. 2004; Hertel et al. 2003; Lakhani and Wolf 2005; Subramanyam and Xia 2008). Comparing with initial participations, there are very few studies that just have begun to explore what mechanisms may sustain long-term voluntary developers' project participations (Fang and Neufeld 2009). Among these studies, Shah (2006) found that long-term participants enjoyed programming and interacting with other developers. This empirical finding suggest that social influence among project members like WOM and OL may play an important role in sustained participations. They also found that initial participations were predominately driven by immediate software use value. Such differences suggest OSS developers may initially join in a project with some short-term needs, but such needs may transform to long-term mechanisms like enjoyment over time. It also implies social influence may have differential impacts on initial and sustained OSS participation.

In a similar vein, Bagozzi and Dholakia (2006) found that sustained participation is associated with developers' senses of identification. Such senses are often strengthened by complex social interactions among project members. Engaged project members view their contribution as “enjoyable joint activities to be done” with their peers. Fang et al. (2009) also found that long-term contributors are influenced by their social interactions with the project community. Von Hippel and von Krogh et al. (2003) found that the momentum for developers’ sustained participation is largely due to their social interactions with other project members. However, all these studies did not examined the impacts of social influence that are embedded on such interactions, which has contributed to strengthened sense of identity and sustained participation.

To summarize, prior OSS participation motivation research focused on the initial participations. The social influence perspective is largely ignored. Meanwhile, the few sustained participation studies have found that social interactions among project members may strengthen developers’ senses of identification and long-term enjoyment, thereby contributing to sustained participations. However, all those studies did not investigate the impacts of social influence.

Based on the above review, we suggested that social influence (WOM and OL) may have differential impacts on initial and sustained OSS participation, mainly due to two reasons. First, the objectivity of the information conveyed through the two social influence mechanisms are rather different. Second, developers’ knowledge level of the underlying project and its members may increase from initial to sustained participation (stage). They may become more familiar with the project and its members and more capable to evaluate these members’ subjective opinions (WOM), thereby can better decide whether to continue to contribute to this project (i.e., sustained participation). However, this needs to be empirically examined. Therefore, we propose the following research questions:

RQ 2.1: What are the impacts of social influence (WOM and OL) on OSS project initial participation?

RQ 2.2: What are the impacts of social influence (WOM and OL) on OSS project sustained participation?

Research Testbed (Data)

The data used in our study was collected from a large online open source software (OSS) development platform/community called Open Hub (formerly known as Ohloh). This platform retrieves OSS development information from major software version control systems. Until 2017, it contains more than 669,000 projects and 3.85 million developers, ranging from well-known projects such like MySQL to lesser-known ones like CakePHP. We developed a set of Java programs to automatically retrieve data through the API of Open Hub. All retrieved data items are in XML format and parsed into a database. Such information includes OSS developers' project participation, their location, nationality, programming language preferences, development activities, and project statistics. Moreover, it also keeps track of the changes in the source code of each listed OSS project in the version control systems and calculated monthly software metrics such as the total number of commits (a commit is a one-time developer's contribution to the source code of an OSS project), total number of developers. The detailed longitudinal development activities at both the project and developer level enable us to construct a panel data sample for our investigation of the research questions

Open Hub provides a microblogging service on its platform from 2008 to 2012. It allows developers to publish information about their opinions, recommendations, and project participation activities through profile web pages of projects and followers, thereby may influence others' participations through WOM. In the profile page of each project, Open Hub displays the project development activity summary in the project profile webpage, such as number of commits (a commit is a one-time developer's contribution to the source code of an OSS project), number of developers. Therefore, developers can also be influenced through OL.

Since the microblogging service was launched in the Open Hub platform in 2008 and it requires developers to first register with the platform to use, the user base is smaller than the millions of developers whose information were automatically collected by Open Hub from major software version systems. Our data sample contains the projects which have developers who use the microblogging service of the Open Hub platform in 2008. We eliminated the projects whose developers do not involve in the microblogging service from our data sample. Table 1 shows the summary statistics of the data sample.

Table 1. Summary Statistics of the Open Hub Data Sample

Time Span of Analysis	Number of Involved Projects	Number of Developers	Number of Microbloggers	Microblogging Messages	Number of Project Tags	Number of Observations
May/2008-February/2012	1716	9432	1839	5795	3302	65,556

Research Methodology

In this section, we present our measurements of OSS project initial and sustained participations, OL, WOM, as well as the empirical model which is used to analyze their relationships.

Dependent Variables

To study the impacts of social influence (WOM and OL) on OSS developers' initial project participation, we use the number of developers who participate in a specific project for the first time in month t ($Monthly\_New\_Participation_{it}$) as the dependent variable. On the other hand, we use the number of new commits to that project in month t ($Monthly\_New\_Commits_{it}$) as the dependent variable for measuring the level of developers' sustained participation.

Independent Variables

Word of Mouth (WOM)

In our empirical analysis, we use the volume of the microblogging messages with the project tag published by existing project members in month t-1 ($Microblogging\_Tags_{i,t\text{-}1}$) as the independent variable for the WOM mechanism. Previous social influence studies usually model both the volume and valence of the WOM messages. We have browsed the Open Hub microblogging messages and found there is little negative content. Most messages are positive opinions or about project progress. Therefore, our model only includes volume of microblogging messages.

Observational Learning (OL)

In the OL process, the actions of prospective participants should be consistent with the actions of existing members they have observed. Thus for the dependent variable initial participation, we use the total number of existing developers in the underlying project in month t-1

($Cumu\_Developers_{i,t-1}$) as the independent variable. Because this number shown in the project profile web page indicates the cumulative number of initial participations since the start of the project. Similarly, we use the cumulative number of commits ($Cumu\_Commits_{i,t-1}$) as the independent variable for sustain participation since it measures the level of cumulative developer contributions in a project.

Control Variables

Cumulative number of code lines *($\boldsymbol{Cumu\_Codelines_{i,t-1}}$):* This variable is often used to measure the complexity of the underlying software project and the project output. (von Hippel and von Krogh 2003) used this measure as an index. We conjecture that a more successful project tends to attract more participations, and we then include it as our control variable.

Project age *($\boldsymbol{Project\_Age_{i,t-1}}$):* Grewal et al. (2006) indicated that project age signals stage of project life cycle. Hahn et al. (2008) suggested that different developers may prefer joining projects at different stage. Subramaniam et al. (2009) argued project age may be a proxy for other factors affecting project success such as the developers' group experience.

$\boldsymbol{Monthly\_New\_Participation_{i,t-1}}$ and $\text{Monthly\_New\_Commits}_{i,t-1}$ are lagged dependent variables and used to control for reverse causality issues. Such issues arises when we cannot distinguish if more social influence (WOM and OL) effects that cause more initial and sustain participations, or more participations have generated more WOM and OL.

Average experience of project members *($Project\_Members\_Exp_{i,t-1}$)* is used in previous study (Roberts et al. 2006) as a proxy for developers' characteristics like knowledge and skills that are difficult to measure and may affect the project performance. We also adopt this measure as a control variable for assessing two of the commonly used OSS project performance measures – the initial and sustained project participations.

Average reputation score of project members ($Project\_Members\_Rep_{i,t-1}$): Our previous research (Hu et al. 2012) has found that developers with good reputation score tend to attract more project collaborators since they may want to learn from these reputable OSS developers in terms of programming or project management. We then adopt average number of project members' reputation as our control variable.

Table 2. Summary of Measures	
Dependent Variables	
$Monthly\_New\_Commits_{it}$	The number of new commits made to the project i in month t.
$Monthly\_New\_Participation_{it}$	The number of new developers who participated in project i in month t.
Independent Variable (WOM)	
$Microblogging\_Tags_{i,t-1}$	The number of microblog messages which contain a tag that links to project i's name in the month t-1.
Independent Variables (OL)	
$Cumu\_Commits_{i,t-1}$ (Sustain Participation)	The cumulative number of commits the project i has until month t-1.
$Cumu\_Developers_{i,t-1}$ (Initial Participation)	The cumulative number of developers the project i has until month t-1.
Control Variables (Current Study)	
$Cumu\_Codelines_{i,t-1}$	The cumulative number of lines of code, excluding comments and blanks of the project i until month t-1.
$Project\_Age_{i,t-1}$	The number of months i existed until month t-1
$Monthly\_New\_Commits_{i,t-1}$	One-month lagged variable of the monthly new commits.
$Monthly\_New\_Participation_{i,t-1}$	One-month lagged variable of the monthly new participation.
Control Variables (Future Study)	
$Project\_Members\_Exp_{i,t-1}$	The average number of months developers spent on the project i until month t-1.
$Project\_Members\_Rep_{i,t-1}$	The average number of the project members' reputation scores (Kudo rank) in the project i.

Panel Regression Model

This study aims to provide a causal interpretation of the observed correlation between the two types of social influence mechanisms and OSS developers' project participation behaviors. We carefully designed our empirical model which leverage the panel structure of our data sample to control for the unobserved heterogeneity in project characteristics and possible endogeneity issues like reverse causality. Again, the dependent variables are the measures of initial and sustained OSS project participation behavior as defined previously, and the WOM and OL measures are used as independents. Controlling for project-level unobserved effects is achieved in the panel model by introducing fixed effects. We also control the project-specific fixed effects ρ_i and η_i in the two models to capture the idiosyncratic characteristics associated with each project, such as project license, programming language, manager etc. In addition, the one-month lagged dependent variables are used in our model for the identification of reverse causality issues. In order to decide between fixed or random effects, we ran a Hausman test. The p-value of the Hausman test results is 0.01207. The p-value is significant (p-value<0.05). Then we choose to use fixed effects in our study. The fixed effects capture the time invariant, unobserved heterogeneity of each project. Thus we can control for unobserved differences across different projects.

$$\text{Monthly\_New\_Participation}_{it} = \alpha_1 \text{Microblogging\_Tags}_{i,t-1} + \alpha_2 \text{Cumu\_Commits}_{i,t-1} + \alpha_3 \text{Cumu\_Developers}_{i,t-1} + \alpha_4 \text{Cumu\_Codelines}_{i,t-1} + \alpha_5 \text{Monthly\_New\_Participation}_{i,t-1} + \alpha_6 \text{Project\_Age}_{i,t-1} + \rho_i + \varepsilon_{it} \qquad (1)$$

$$\text{Monthly\_New\_Commits}_{it} = \beta_1 \text{Microblogging\_Tags}_{i,t-1} + \beta_2 \text{Cumu\_Commits}_{i,t-1} + \beta_3 \text{Cumu\_Developers}_{i,t-1} + \beta_4 \text{Cumu\_Codelines}_{i,t-1} + \beta_5 \text{Monthly\_New\_Commits}_{i,t-1} + \beta_6 \text{Project\_Age}_{i,t-1} + \eta_i + \sigma_{it} \qquad (2)$$

Preliminary Results and Discussions

Table 3. Preliminary Results

Variables	Initial Participation	Sustained Participation
$Microblogging\_Tags_{i,t-1}$	.0125 (1.70e-03)***	-.0223 (1.70e-03)
$Cumu\_Commits_{i,t-1}$	.0275 (4.72e-03)***	.0026 (1.04e-04)***
$Cumu\_Developers_{i,t-1}$	.0566 (3.43e-03)***	.0652 (5.68e-03)***
$Cumu\_Codelines_{i,t-1}$	.0247 (2.19e-03)***	.0001 (7.92e-07)***
$Monthly\_New\_Developers_{i,t-1}$	.2701 (3.79e-03)***	-
$Monthly\_New\_Commits_{i,t-1}$	-	.6992 (2.73e-03)***
$Project\_Age_{i,t-1}$	-.1429 (4.78e-03)***	-.4559 (5.01e-02)***
Number of Observations	65,556	65,556
R^2	0.41	0.43

*** p<0.01, ** p<0.05, *p<0.1. Standard errors are reported in parentheses.

Table 3 shows the preliminary results of our panel data analysis. It was found that the impacts of WOM on OSS developers' initial and sustained project participation differ from each other. The coefficient of the $Microblogging\_Tags_{i,t-1}$ is positive and significant in model (1) but not in model (2). This means that project i's prospect participants are significantly influenced by i's related microblogging messages for their initial participation. As these participants become more familiar with the project and its members after initial participation, such WOM influence disappears. As mentioned before, we conjecture it is because that project and member familiarity enable developers to better evaluate the more subjective information conveyed through other members' WOM.

On the other hand, OL effects existed for both OSS developers' initial and sustained participation. This may be because that the OL information is often objective statistics. Its impacts are difficult to change when developers' own familiarity or perception of the underlying project changes in the stage of sustained participation.

Moreover, the coefficient of the $Project\_Age_{i,t-1}$ is negative and significant in both model (1) and (2). This is consistent with prior OSS studies and indicates that older projects are less likely to attract initial and sustained participation from OSS developers. $Cumu\_Commits_{i,t-1}$ represents the cumulative number of lines of code added to the OSS project. Prior OSS project participation study(Fang and Neufeld 2009; Roberts et al. 2006) also argued that it can influence the developers' participation behaviors. Our preliminary results support this argument as well. In order to alleviate the concern of the reverse causality in our study, we presented the results of the one-month lag of two dependent variables.

These results may provide important practical implications for OSS stakeholders to use social influence to manage OSS participation. Major OSS platforms like Open Hub and GitHub has adopted various social networking functions like microblogging and reputation systems for a long time. First, OSS project managers can leverage resources to encourage more positive microblogging (WOM) messages and publish detailed OSS participation activities (OL) in their project profile pages. However, when they aim to attract more sustained participation, it may be better to shift more resources from WOM to OL based methods. Second, based on results of our future analysis, we would like to know if publishing more experienced or highly recognized developers' participation activities can improve initial and sustained participation respectively. Third, we also would like to explore if WOM and OL mechanisms can work together and better improve OSS participation together.

Limitations, Future Work, and Intended Contributions

Since this is an ongoing research and we have not fully finished data processing of all proposed variables, the empirical model used in our preliminary analysis is incomplete. We may miss control variables that can be major drivers of initial and sustained contributions. This could affect our current preliminary results. We will keep improving the model and finish the full analyses soon.

For this ongoing study and future work, we first would like to improve our model by adding more control variables based on the OSS and social influence literature and our empirical setting. For instance, as shown in the last two rows of table 2, we will examine if having more experienced or highly recognized developers can help an OSS project to attract more initial and sustained participation, and how they may affect the impacts of WOM and OL. Second, we would like to examine the interaction effects of WOM and OL to see if they complement or compete with each other in terms of improving OSS participation. Third, we would like to conduct sub group analysis to find out if project characteristics such as project age, size, and structure, may moderate the impacts of social influence. Furthermore, in order to make the contributions of our findings more solid and creditable, we will also conduct our study with other existing methodology or other open source software data sets.

We intend to make three contributions. First, to the best of our knowledge, we are among the first to adopt a social influence perspective to study OSS developers' initial and sustained participation behaviors. It can improve our theoretical understanding of the impacts of social influence on individuals' behaviors that are driven by economic goals. Second, we studied the possible differential impacts of the WOM and OL on OSS participations. Third, we investigated how social influence may affect the initial OSS participation and the sustained participation. Moreover, we hope our analysis can bring us empirical insights that can help design better social functions and strategies to attract more initial and sustained OSS project participation

References

Allen, T. J. 1977. "Managing the Flow of Technology: Technology Transfer and the Dissemination of Technological Information within the R and D Organization,").

Aral, S. 2011. "Commentary—Identifying Social Influence: A Comment on Opinion Leadership and Social Contagion in New Product Diffusion," *Marketing Science* (30:2), pp. 217-223.

Aral, S., Muchnik, L., and Sundararajan, A. 2009. "Distinguishing Influence-Based Contagion from Homophily-Driven Diffusion in Dynamic Networks," *Proceedings of the National Academy of Sciences* (106:51), pp. 21544-21549.

Aral, S., and Walker, D. 2011. "Creating Social Contagion through Viral Product Design: A Randomized Trial of Peer Influence in Networks," *Management science* (57:9), pp. 1623-1639.

Aral, S., and Walker, D. 2012. "Identifying Influential and Susceptible Members of Social Networks," *Science* (337:6092), pp. 337-341.

Arndt, J. 1967. "Role of Product-Related Conversations in the Diffusion of a New Product," *Journal of marketing Research*), pp. 291-295.

Awad, N. F., and Ragowsky, A. 2008. "Establishing Trust in Electronic Commerce through Online Word of Mouth: An Examination across Genders," *Journal of Management Information Systems* (24:4).

Bagozzi, R. P., and Dholakia, U. M. 2006. "Open Source Software User Communities: A Study of Participation in Linux User Groups," *Management Science* (52:7), pp. 1099-1115.

Bandura, A. 1971. "Social Learning Theory. Morristown." NJ: General Learning Press.

Bandura, A. 1977. "Social Learning Theory Englewood Cliffs." NJ: Prentice-Hall.

Banerjee, A. 1992. "A Simple Model of Herd Behavior," *Quarterly Journal of Economics* (107:3), pp. 797-817.

Barbagallo, D., Francalenei, C., and Merlo, F. 2008. "The Impact of Social Netowrking on Software Design Quality and Development Effort in Open Source Projects," *ICIS 2008 proceedings*), p. 201.

Bell, D. R., and Song, S. 2007. "Neighborhood Effects and Trial on the Internet: Evidence from Online Grocery Retailing," *Quantitative Marketing and Economics* (5:4), pp. 361-400.

Bergiel, B. J., Bergiel, E. B., and Balsmeier, P. W. 2006. "The Reality of Virtual Teams," *Competition Forum*: American Society for Competitiveness, p. 427.

Bikhchandani, S., Hirshleifer, D., and Welch, I. 1992. "Learning from the Behavior of Others: Conformity, Fads, and Information Cascades," *Journal of Political Perspectives* (12:3), pp. 151-170.

Branson, L., Clausen, T. S., and Sung, C.-H. 2008. "Group Style Differences between Virtual and F2f Teams," *American Journal of Business* (23:1), pp. 65-70.

Burke, K., Aytes, K., Chidambaram, L., and Johnson, J. J. 1999. "A Study of Partially Distributed Work Groups the Impact of Media, Location, and Time on Perceptions and Performance," *Small Group Research* (30:4), pp. 453-490.

Burke, M., Marlow, C., and Lento, T. 2009. "Feed Me: Motivating Newcomer Contribution in Social Network Sites," *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, Boston, MA: ACM.

Chatzisarantis, N. L., Hagger, M. S., and Smith, B. 2007. "Influences of Perceived Autonomy Support on Physical Activity within the Theory of Planned Behavior," *European Journal of Social Psychology* (37:5), pp. 934-954.

Chen, Y., Wang, Q., and Xie, J. 2011. "Online Social Interactions: A Natural Experiment on Word of Mouth Versus Observational Learning," *Journal of marketing research* (48:2), pp. 238-254.

Chen, Y., and Xie, J. 2008. "Online Consumer Review: Word-of-Mouth as a New Element of Marketing Communication Mix," *Management Science* (54:3), pp. 477-491.

Cheung, M. K. C., Liu, L. B. I., and Lee, K. O. M. 2015. "How Online Social Interactions Influence Customer Information Contribution Behavior in Online Social Shopping Communities: A Social Learning Theory Perspective," *Journal of the Association for Information Science and Technology* (66:12), pp. 2511-2521.

Cheung, M. K. C., and Thadani, R. D. 2012. "The Impact of Electronic Word-of-Mouth Communication: A Literature Analysis and Integrative Model," *Decision Support Systems* (54:1), pp. 461-470.

Cheung, M. K. C., Xiao, B. S., and Liu, L. B. I. 2014. "Do Actions Speak Louder Than Voices? The Signaling Role of Social Information Cues in Influencing Consumer Puchase Decisions," *Decision Support Systems* (65), pp. 50-58.

Chevalier, J. A., and Mayzlin, D. 2006. "The Effect of Word of Mouth on Sales: Online Book Reviews," *Journal of Marketing Research* (43:9), pp. 345-354.

Choi, J., Hui, S. K., and Bell, D. R. 2010. "Spatiotemporal Analysis of Imitation Behavior across New Buyers at an Online Grocery Retailer," *Journal of Marketing Research* (47:1), pp. 75-89.

Chudoba, K. M., Wynn, E., Lu, M., and Watson-Manheim, M. B. 2005. "How Virtual Are We? Measuring Virtuality and Understanding Its Impact in a Global Organization," *Information systems journal* (15:4), pp. 279-306.

Clemens, E., Gao, G., and Hitt, L. 2006. "When Online Reviews Meet Hyperdifferentiation: A Study of the Craft Beer Industry," *Journal of Management Information Systems* (23:2), pp. 149-171.

Colazo, J. A., and Fang, Y. 2010. "Following the Sun: Temporal Dispersion and Performance in Open Source Software Project Teams," *Journal of the Association for Information Systems* (11:11), p. 684.

Cramton, C. D. 2001. "The Mutual Knowledge Problem and Its Consequences for Dispersed Collaboration," *Organization science* (12:3), pp. 346-371.

Crowston, K., Annabi, H., and Howison, J. 2003. "Defining Open Source Software Project Success," *ICIS 2003 Proceedings*), p. 28.

Crowston, K., Howison, J., and Annabi, H. 2006. "Information Systems Success in Free and Open Source Software Development: Theory and Measures," *Software Process: Improvement and Practice* (11:2), pp. 123-148.

Crowston, K., Howison, J., Masango, C., and Eseryel, U. Y. 2005. "Face-to-Face Interactions in Self-Organizing Distributed Teams," *Academy of Management Conference, Honolulu, HI*.

Crowston, K., Howison, J., Masango, C., and Eseryel, U. Y. 2007. "The Role of Face-to-Face Meetings in Technology-Supported Self-Organizing Distributed Teams," *IEEE Transactions on professional communication* (50:3), pp. 185-203.

Crowston, K., and Scozzi, B. 2002. "Open Source Software Projects as Virtual Organisations: Competency Rallying for Software Development," *IEE Proceedings-Software* (149:1), pp. 3-17.

Dabbish, L., Stuart, C., Tsay, J., and Herbsleb, J. 2012. "Social Coding in Github: Transparency and Collaboration in an Open Software Repository," *Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work*: ACM, pp. 1277-1286.

Daft, R. L., and Lengel, R. H. 1986. "Organizational Information Requirements, Media Richness and Structural Design," *Management science* (32:5), pp. 554-571.

Dahlander, L., and O'Mahony, S. 2011. "Progressing to the Center: Coordinating Project Work," *Organization science* (22:4), pp. 961-979.

Daniel, S., Agarwal, R., and Stewart, K. J. 2013. "The Effects of Diversity in Global, Distributed Collectives: A Study of Open Source Project Success," *Information Systems Research* (24:2), pp. 312-333.

Davidson, E., and Vaast, E. 2009. "Tech Talk: An Investigation of Blogging in Technology Innovation Discourse," *IEEE Transactions on professional communication* (52:1), pp. 40-60.

Davison, R., Bélanger, F., Ahuja, M., Beth Watson-Manheim, M., Espinosa, J. A., DeLone, W., and Lee, G. 2006. "Global Boundaries, Task Processes and Is Project Success: A Field Study," *Information Technology & People* (19:4), pp. 345-370.

Deci, E. L., and Ryan, R. M. 1985. "The General Causality Orientations Scale: Self-Determination in Personality," *Journal of research in personality* (19:2), pp. 109-134.

Dellarocas, C. 2003. "The Digitization of Word-of-Mouth: Promise and Challenges of Online Feeback Machanisms," *Management Science* (49:10), pp. 1407-1424.

Dellarocas, C., Gao, G., and Narayan, R. 2010. "Are Consumers More Likely to Contribute Online Reviews for Hit or Nich Products?," *Journal of Management Information Systems* (27:2), pp. 127-158.

Dennis, A. R., Fuller, R. M., and Valacich, J. S. 2008. "Media, Tasks, and Communication Processes: A Theory of Media Synchronicity," *MIS quarterly* (32:3), pp. 575-600.

Dennis, A. R., George, J. F., Jessup, L. M., Nunamaker Jr, J. F., and Vogel, D. R. 1988. "Information Technology to Support Electronic Meetings," *MIS quarterly*), pp. 591-624.

Dennis, A. R., and Kinney, S. T. 1998. "Testing Media Richness Theory in the New Media: The Effects of Cues, Feedback, and Task Equivocality," *Information systems research* (9:3), pp. 256-274.

Dennis, A. R., and Valacich, J. S. 1999. "Rethinking Media Richness: Towards a Theory of Media Synchronicity," *Systems Sciences, 1999. HICSS-32. Proceedings of the 32nd Annual Hawaii International Conference on*: IEEE, p. 10 pp.

Di Gangi, P. M., and Wasko, M. 2009. "Steal My Idea! Organizational Adoption of User Innovations from a User Innovation Community: A Case Study of Dell Ideastorm," *Decision Support Systems* (48:1), pp. 303-312.

Dietz, R. D. 2002. "The Estimation of Neighborhood Effects in the Social Sciences: An Interdisciplinary Approach," *Social science research* (31:4), pp. 539-575.

Dodds, P. S., and Watts, D. J. 2004. "Universal Behavior in a Generalized Model of Contagion," *Physical review letters* (92:21), p. 218701.

Duan, W., Gu, B., and Whinston, A. B. 2008. "The Dynamics of Online Word-of-Mouth and Product Sales - an Empirical Investigation of the Movie Industry," *Journal of Retailing* (84:2), pp. 233-242.

Dubé, L., and Paré, G. 2004. "The Multifaceted Nature of Virtual Teams," *Virtual teams: Projects, protocols and processes*), pp. 1-39.

Ehrlich, K., Helander, M., Valetto, G., Davies, S., and Williams, C. 2008. "1an Analysis of Congruence Gaps and Their Effect on Distributed Software Development,").

Espinosa, J. A., and Carmel, E. 2003. "The Impact of Time Separation on Coordination in Global Software Teams: A Conceptual Foundation," *Software Process: Improvement and Practice* (8:4), pp. 249-266.

Fang, Y., and Neufeld, D. 2009. "Understanding Sustained Participation in Open Source Software Projects," *Journal of Management Information Systems* (25:4), pp. 9-50.

Festinger, L., Back, K. W., and Schachter, S. 1950. *Social Pressures in Informal Groups: A Study of Human Factors in Housing*. Stanford University Press.

Franke, N., and Von Hippel, E. 2003. "Satisfying Heterogeneous User Needs Via Innovation Toolkits: The Case of Apache Security Software," *Research policy* (32:7), pp. 1199-1215.

Gallivan, M. J. 2001. "Striking a Balance between Trust and Control in a Virtual Organization: A Content Analysis of Open Source Software Case Studies," *Information Systems Journal* (11:4), pp. 277-304.

Ghosh, R. A. 2005. "Understanding Free Software Developers: Findings from the Floss Study," *Perspectives on free and open source software*), pp. 23-46.

Goffman, E. 2005. *Interaction Ritual: Essays in Face to Face Behavior*. AldineTransaction.

Grewal, R., Lilien, G. L., and Mallapragada, G. 2006. "Location, Location, Location: How Network Embeddedness Affects Project Success in Open Source Systems," *Management Science* (52:7), pp. 1043-1056.

Griffith, T. L., Sawyer, J. E., and Neale, M. A. 2003. "Virtualness and Knowledge in Teams: Managing the Love Triangle of Organizations, Individuals, and Information Technology," *MIS quarterly*), pp. 265-287.

Grinter, R. E., Herbsleb, J. D., and Perry, D. E. 1999. "The Geography of Coordination: Dealing with Distance in R&D Work," *Proceedings of the international ACM SIGGROUP conference on Supporting group work*: ACM, pp. 306-315.

Hagerstrand, T. 1967. "Innovation Diffusion as a Spatial Process," *Innovation diffusion as a spatial process.*).

Hahn, J., Moon, J. Y., and Zhang, C. 2008. "Emergence of New Project Teams from Open Source Software Developer Networks: Impact of Prior Collaboration Ties " *Information Systems Research* (19:3), pp. 369-391.

Hambley, L. A., O'Neill, T. A., and Kline, T. J. 2007. "Virtual Team Leadership: The Effects of Leadership Style and Communication Medium on Team Interaction Styles and Outcomes," *Organizational behavior and human decision processes* (103:1), pp. 1-20.

Hambrick, D. C., Davison, S. C., Snell, S. A., and Snow, C. C. 1998. "When Groups Consist of Multiple Nationalities: Towards a New Understanding of the Implications," *Organization studies* (19:2), pp. 181-205.

Hampton, K., and Wellman, B. 2003. "Neighboring in Netville: How the Internet Supports Community and Social Capital in a Wired Suburb," *City & Community* (2:4), pp. 277-311.

Hampton, K. N. 2007. "Neighborhoods in the Network Society the E-Neighbors Study," *Information, Communication & Society* (10:5), pp. 714-748.

Hann, I.-H., Roberts, J., and Slaughter, S. 2004. "Why Developers Participate in Open Source Software Projects: An Empirical Investigation," *ICIS 2004 Proceedings*), p. 66.

Harrison, D. A., and Klein, K. J. 2007. "What's the Difference? Diversity Constructs as Separation, Variety, or Disparity in Organizations," *Academy of management review* (32:4), pp. 1199-1228.

Hars, A., and Ou, S. 2001. "Working for Free? Motivations of Participating in Open Source Projects," *System Sciences, 2001. Proceedings of the 34th Annual Hawaii International Conference on*: IEEE, p. 9 pp.

Hart, R. 2001. "Constituting Relationships in Communication: An Interdisciplinary Approach to Understanding Peer Relationships in Geographically Dispersed Teams," *Advances in Interdisciplinary Studies of Work Teams* (8), pp. 85-106.

Hartmann, W. R., Manchanda, P., Nair, H., Bothner, M., Dodds, P., Godes, D., Hosanagar, K., and Tucker, C. 2008. "Modeling Social Interactions: Identification, Empirical Methods and Policy Implications," *Marketing letters* (19:3-4), pp. 287-304.

Hemetsberger, A. 2002. "Fostering Cooperation on the Internet: Social Exchange Processes in Innovative Virtual Consumer Communities," *NA-Advances in Consumer Research Volume 29*).

Hennig-Thusrau, T., Wiertz, C., and Feldharus, F. 2015. "Does Twitter Matter? The Impact of Microblogging Word of Mouth Con Consumers' Adoption of New Movies," *Journal of the Academy of Marketing Science* (43:3), pp. 375-394.

Herbsleb, J. D., and Mockus, A. 2003. "An Empirical Study of Speed and Communication in Globally Distributed Software Development," *IEEE Transactions on software engineering* (29:6), pp. 481-494.

Herbsleb, J. D., Mockus, A., Finholt, T. A., and Grinter, R. E. 2000. "Distance, Dependencies, and Delay in a Global Collaboration," *Proceedings of the 2000 ACM conference on Computer supported cooperative work*: ACM, pp. 319-328.

Hertel, G., Niedner, S., and Herrmann, S. 2003. "Motivation of Software Developers in Open Source Projects: An Internet-Based Survey of Contributors to the Linux Kernel," *Research policy* (32:7), pp. 1159-1177.

Hippel, E. v., and Krogh, G. v. 2003. "Open Source Software and the “Private-Collective” Innovation Model: Issues for Organization Science," *Organization science* (14:2), pp. 209-223.

Hu, D., Zhao, J. L., and Cheng, J. 2012. "Reputation Management in an Open Source Developer Social Network: An Empirical Study on Determinants of Positive Evaluations," *Decision Support Systems* (53:3), pp. 526-533.

Jalali, S., and Wohlin, C. 2012. "Global Software Engineering and Agile Practices: A Systematic Review," *Journal of software: Evolution and Process* (24:6), pp. 643-659.

Jannik Meyners, C. B., Jan U. Becker, Jacob Goldenberg. 2017. "The Role of Mere Closeness: How Geographic Proximity Affects Social Influence," *Journal of Marketing*).

Jansen, B. J., Zhang, M., Sobel, K., and Chowdury, A. 2009. "Twitter Power: Tweets as Electronic Word of Mouth," *Journal of the Association for Information Science and Technology* (60:11), pp. 2169-2188.

Java, A., Song, X., Finin, T., and Tseng, B. 2007. "Why We Twitter: Understanding Microblogging Usage and Communities," *Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis*: ACM, pp. 56-65.

Joshi, A., and Roh, H. 2009. "The Role of Context in Work Team Diversity Research: A Meta-Analytic Review," *Academy of Management Journal* (52:3), pp. 599-627.

Katz, E., and Lazarsfeld, P. F. 1966. *Personal Influence, the Part Played by People in the Flow of Mass Communications*. Transaction Publishers.

Kiesler, S., and Cummings, J. N. 2002. "What Do We Know About Proximity and Distance in Work Groups? A Legacy of Research," *Distributed work* (1), pp. 57-80.

Kraut, R. E., and Streeter, L. A. 1995. "Coordination in Software Development," *Communications of the ACM* (38:3), pp. 69-82.

Krishnamurthy, S. 2003. "A Managerial Overview of Open Source Software," *Business Horizons* (46:5), pp. 47-56.

Krishnamurthy, S. 2006. "On the Intrinsic and Extrinsic Motivation of Free/Libre/Open Source (Floss) Developers," *Knowledge, Technology & Policy* (18:4), pp. 17-39.

Krogh, G. V., Haefliger, S., Spaeth, S., and Wallin, M. W. 2012. "Carrots and Rainbows: Motivation and Social Practice in Open Source Software Development," *MIS Q.* (36:2), pp. 649-676.

Lakhani, K. R., and Wolf, R. G. 2005. "Why Hackers Do What They Do: Understanding Motivation and Effort in Free/Open Source Software Projects," *Perspectives on free and open source software* (1), pp. 3-22.

Lau, D. C., and Murnighan, J. K. 1998. "Demographic Diversity and Faultlines: The Compositional Dynamics of Organizational Groups," *Academy of Management Review* (23:2), pp. 325-340.

Lerner, J., and Tirole, J. 2002. "Some Simple Economics of Open Source," *The journal of industrial economics* (50:2), pp. 197-234.

Levy, M., and Goldenberg, J. 2014. "The Gravitational Law of Social Interaction," *Physica A: Statistical Mechanics and its Applications* (393), pp. 418-426.

Lipnack, J., and Stamps, J. 1999. "Virtual Teams: The New Way to Work," *Strategy & Leadership* (27:1), pp. 14-19.

Luthiger, B., and Jungwirth, C. 2007. "Pervasive Fun," *First Monday* (12:1).

Malone, T. W., and Crowston, K. 1994. "The Interdisciplinary Study of Coordination," *ACM Computing Surveys (CSUR)* (26:1), pp. 87-119.

Martins, L. L., Gilson, L. L., and Maynard, M. T. 2004. "Virtual Teams: What Do We Know and Where Do We Go from Here?," *Journal of management* (30:6), pp. 805-835.

Mihhailova, G. 2007. "Virtual Teams: Just a Theoretical Concept or a Widely Used Practice," *The Business Review, Cambridge* (7:1), pp. 186-193.

Mizerski, R. 1982. "An Attribution Explaination of the Disproportionate Influence of Unfavorable Information," *Journal of Consumer Research* (9:3), pp. 301-310.

Nguyen, T., Wolf, T., and Damian, D. 2008. "Global Software Development and Delay: Does Distance Still Matter?," *Global Software Engineering, 2008. ICGSE 2008. IEEE International Conference on*: IEEE, pp. 45-54.

O'Leary, M. B., and Cummings, J. N. 2007. "The Spatial, Temporal, and Configurational Characteristics of Geographic Dispersion in Teams," *MIS quarterly* (31:3), pp. 433-452.

O'Leary, M. B., and Mortensen, M. 2010. "Go (Con) Figure: Subgroups, Imbalance, and Isolates in Geographically Dispersed Teams," *Organization Science* (21:1), pp. 115-131.

Olson, J. S., Teasley, S., Covi, L., and Olson, G. 2002. "The (Currently) Unique Advantages of Collocated Work," *Distributed work*), pp. 113-135.

Purvanova, R. K., and Bono, J. E. 2009. "Transformational Leadership in Context: Face-to-Face and Virtual Teams," *The Leadership Quarterly* (20:3), pp. 343-357.

Ramasubbu, N., Cataldo, M., Balan, R. K., and Herbsleb, J. D. 2011. "Configuring Global Software Teams: A Multi-Company Analysis of Project Productivity, Quality, and Profits," *Proceedings of the 33rd international conference on Software engineering*: ACM, pp. 261-270.

Raymond, E. 1999. "The Cathedral and the Bazaar: Musings on Linux and Open Source from an Accidental Revolutionary. Sebastapol," *Cal.: O'Reilly*).

Raymond, E. S. 2001. *The Cathedral & the Bazaar: Musings on Linux and Open Source by an Accidental Revolutionary*. " O'Reilly Media, Inc.".

Roberts, J. A., Hann, I.-H., and Slaughter, S. A. 2006. "Understanding the Motivations, Participation, and Performance of Open Source Software Developers: A Longitudinal Study of the Apache Projects," *Management science* (52:7), pp. 984-999.

Rui, H., Liu, Y., and Whinston, A. B. 2013. "Whose and What Chatter Matters? The Effect of Tweets on Movie Sales," *Decision Support Systems* (55:4), pp. 863-870.

Santos, C., Kuk, G., Kon, F., and Pearson, J. 2013. "The Attraction of Contributors in Free and Open Source Software Projects," *Journal of Strategic Information Systems* (22:1), pp. 26-45.

Scacchi, W., Feller, J., Fitzgerald, B., Hissam, S., and Lakhani, K. 2006. "Understanding Free/Open Source Software Development Processes," *Software Process: Improvement and Practice* (11:2), pp. 95-105.

Seebach, C., Beck, R., and Pahlke, I. 2011. "Situation Awareness through Social Collaboration Platforms in Distributed Work Environments,").

Sessions, L. F. 2010. "How Offline Gatherings Affect Online Communities: When Virtual Community Members 'Meetup'," *Information, Communication & Society* (13:3), pp. 375-395.

Shah, S. K. 2006. "Motivation, Governance, and the Viability of Hybrid Forms in Open Source Software Development," *Management science* (52:7), pp. 1000-1014.

Sharp, H., Giuffrida, R., and Melnik, G. 2012. "Information Flow within a Dispersed Agile Team: A Distributed Cognition Perspective," *International Conference on Agile Software Development*: Springer, pp. 62-76.

Singh, P. V., and Phelps, C. 2012. "Networks, Social Influence, and the Choice among Competing Innovations: Insights from Open Source Software Licenses," *Information Systems Research* (24:3), pp. 539-560.

Šmite, D., Wohlin, C., Gorschek, T., and Feldt, R. 2010. "Empirical Evidence in Global Software Engineering: A Systematic Review," *Empirical software engineering* (15:1), pp. 91-118.

Staples, D. S., and Webster, J. 2008. "Exploring the Effects of Trust, Task Interdependence and Virtualness on Knowledge Sharing in Teams," *Information Systems Journal* (18:6), pp. 617-640.

Staples, D. S., and Zhao, L. 2006. "The Effects of Cultural Diversity in Virtual Teams Versus Face-to-Face Teams," *Group Decision and Negotiation* (15:4), pp. 389-406.

Stewart, D. 2005. "Social Status in an Open-Source Community," *American Sociological Review* (70:5), pp. 823-842.

Stewart, K. J., Ammeter, A. P., and Maruping, L. M. 2006. "Impacts of License Choice and Organizational Sponsorship on User Interest and Development Activity in Open Source Software Projects," *Information Systems Research* (17:2), pp. 126-144.

Subramaniam, C., Sen, R., and Nelson, M. L. 2009. "Determinants of Open Source Software Project Success: A Longitudinal Study," *Decision Support Systems* (46:2), pp. 576-585.

Subramanyam, R., and Xia, M. 2008. "Free/Libre Open Source Software Development in Developing and Developed Countries: A Conceptual Framework with an Exploratory Study," *Decision support systems* (46:1), pp. 173-186.

Te"eni, D. 2001. "Review: A Cognitive-Affective Model of Organizational Communication for Designing It," *MIS quarterly* (25:2), pp. 251-312.

Townsend, A. M., DeMarie, S. M., and Hendrickson, A. R. 1998. "Virtual Teams: Technology and the Workplace of the Future," *The Academy of Management Executive* (12:3), pp. 17-29.

Tsay, J. T., Dabbish, L., and Herbsleb, J. 2012. "Social Media and Success in Open Source Projects," *Proceedings of the ACM 2012 conference on computer supported cooperative work companion*: ACM, pp. 223-226.

Van den Bulte, C., and Moenaert, R. K. 1998. "The Effects of R&D Team Co-Location on Communication Patterns among R&D, Marketing, and Manufacturing," *Management Science* (44:11-part-2), pp. S1-S18.

Van Knippenberg, D., and van Ginkel, W. P. 2010. "The Categorization-Elaboration Model of Work Group Diversity: Wielding the Double-Edged Sword,").

von Hippel, E., and von Krogh, G. 2003. "Open Source Software and the "Private-Collective" Innovation Model: Issues for Organization Science," *Organization Science* (14:2), pp. 209-223.

Von Krogh, G., Haefliger, S., Spaeth, S., and Wallin, M. W. 2012. "Carrots and Rainbows: Motivation and Social Practice in Open Source Software Development," *MIS quarterly* (36:2).

Von Krogh, G., Spaeth, S., and Lakhani, K. R. 2003. "Community, Joining, and Specialization in Open Source Software Innovation: A Case Study," *Research Policy* (32:7), pp. 1217-1241.

Wakefield, R. L., Leidner, D. E., and Garrison, G. 2008. "Research Note-a Model of Conflict, Leadership, and Performance in Virtual Teams," *Information Systems Research* (19:4), pp. 434-455.

Warkentin, M. E., Sayeed, L., and Hightower, R. 1997. "Virtual Teams Versus Face-to-Face Teams: An Exploratory Study of a Web-Based Conference System," *Decision Sciences* (28:4), pp. 975-996.

Williams, E. 1977. "Experimental Comparisons of Face-to-Face and Mediated Communication: A Review," *Psychological Bulletin* (84:5), p. 963.

Yang, S., and Allenby, G. M. 2003. "Modeling Interdependent Consumer Preferences," *Journal of Marketing Research* (40:3), pp. 282-294.

Ye, Q., Cheng, and Fang, B. 2013. "Learning from Other Buyers: The Effect of Purchase History Records in Online Marketplaces," *Decision Support Systems* (56), pp. 502-512.

Yoshikawa, Y., Iwata, T., and Sawada, H. 2014. "Collaboration on Social Media: Analyzing Successful Projects on Social Coding," *arXiv preprint arXiv:1408.6012*).

Zhao, D., and Rosson, M. B. 2009. "How and Why People Twitter: The Role That Micro-Blogging Plays in Informal Communication at Work," *Proceedings of the ACM 2009 international conference on Supporting group work*: ACM, pp. 243-252.

Zhou, Z., Wu, J. P., Zhang, Q., and Xu, S. 2013. "Transforming Visitors into Members in Online Brand Communities: Evidence from China," *Journal of Business Research* (66:12), pp. 2438-2443.

Curriculum Vitae

Xiao Li

E-mail: davidlixiao@icloud.com Tel: +41 79 7288 159

Education	Ph.D. in Computer Science Department of Informatics, University of Zurich Supervisor: Prof. Dr. Daning Hu	2014-2018 Switzerland
	M.Sc. in Software Engineering University of Science and Technology of China	2010-2013 Hefei China
	B.Sc. in Software Engineering DaLian JiaoTong University	2004-2009 Dalian China
Research Experience	Research Assistant Business Intelligence Research Group, Department of Informatics University of Zurich	2014-2018
	Research Intern Student State Key Laboratory of Computer Science, Institute of Software Chinese Academy of Sciences	2011-2013
Research Interests	Causal Inference (Econometric Identification Methods), Bayesian Data Analysis, Deep Learning and Statistical Learning, Social Computing, Business Intelligence.	
Language Skills	Chinese: Native English: Professional working proficiency Japanese: Basic	
Programming Skills	R/Python/Java SQL Server/MySQL	
Awards	Scholarship of Dalian JiaoTong University Scholarship of Dalian JiaoTong University Excellent Student of Dalian JiaoTong University	2006-2007 2007-2008 2007-2009
Certification	[1]: Neural Networks and Deep Learning (Coursera) [2]: Improving Deep Neural Networks (Coursera) [3]: Structuring Machine Learning Projects (Coursera)	

List of Publications

[1]: Understanding the Impacts of Social Influence on Initial and Sustained Participation in Open Source Software Projects, In: International Conference on Information Systems, Seoul, Korea, 2017 (With Xuan Yang, Daning Hu, Ji Wu, Jiannan Wang)

[2]: The Impacts of Geographic Dispersion on OSS Project Success: Face-to-face vs. Virtual Collaboration, In: International Conference on Information Systems, Dublin, Ireland, 2016 (With Daning Hu, Xiaoquan Zhang)

[3]: Systemic risk management and investment analysis with financial network analytics: research opportunities and challenges, Financial Innovation, Vol. 1 (2), 2015. (With Daning Hu, Gerhard Schwabe)